Institut für Logik, Komplexität und Deduktionssysteme
der Universität Karlsruhe
Lehrstuhl Prof. A. Waibel

Akustische Modellierung sprachlicher und nichtsprachlicher Geräusche

Studienarbeit
von

Tanja Schultz
(tanja@ira.uka.de)
Betreuer: Ivica Rogina

Karlsruhe, den 22. Juni 1994

Zusammenfassung

Jüngster Forschungsgegenstand in der Spracherkennung ist spontan gesprochene Sprache. Charakteristisches Phänomen spontaner Sprache ist der mangelnde Sprachfluß, der unter anderem durch Sprechpausen bedingt wird, die mit verschiedenartigen Geräuschen gefüllt sein können.

Ziel dieser Arbeit war die Verbesserung der Erkennungsleistung von JANUS auf spontaner Sprache durch eine geeignete akustische Modellierung von sprachlichen und nichtsprachlichen Geräuschen.

Derzeitiges Hauptproblem beim Arbeiten mit Geräuschen ist das geringe Trainingsmaterial. Es wurde untersucht, ob die Konzentration des Trainingsmaterials auf wenige Geräuschklassen zu einer Verbesserung der Erkennungsleistung führt. Dazu wurde eine Clusterung der verschiedenen Geräuschklassen vorgenommen und die Ergebnisse der verschiedenen Clustervarianten miteinander verglichen.

Inhaltsverzeichnis

Tabellenverzeichnis

Abbildungsverzeichnis

1 Einleitung und Problemstellung

In jüngerer Zeit beschäftigt sich die Forschung auf dem Gebiet der Spracherkennung mit dem Problem spontan gesprochener Sprache. Spontane Sprache wirft gegenüber gelesener Sprache viele neue, zusätzliche Probleme auf. Untergliedert man die Gesamtaufgabe „Spracherkennung“ in die drei Teilbereiche Akustik – Syntax – Semantik, dann entstehen bei spontaner Sprache im Einzelnen zusätzlich:

- Probleme auf der akustisch–phonetischen Ebene durch
  - starke Variation von Geschwindigkeit und Artikulation, sowohl Satz– als auch Sprecherübergreifend betrachtet
  - sehr unterschiedlich lange Sätze und ausgeprägte Prosodie
  - wenig Sprachfluß, d.h. unflüssig gesprochene Sätze
- Probleme auf der syntaktisch–grammatikalischen Ebene durch
  - nur bedingte Kontrolle über lexikalische Inhalte, d.h. offener Sprachschatz, unbekannte Wörter
  - keine Kontrolle über syntaktische Inhalte, d.h. grammatikalisch falsche Satzkonstruktionen
- Probleme auf der semantisch–inhaltlichen Ebene durch
  - freie Variationsmöglichkeiten der Sprecher, die die Interpretation erschweren
  - häufigeres Auftreten von Zweideutigkeiten

Die vorliegende Studienarbeit beschäftigt sich ausschließlich mit Problemen auf akustischer Ebene und hier mit dem Phänomen der mangelnden Flüssigkeit spontan gesprochener Sprache. Unterbrechungen im Sprachfluß sind charakteristisch für die spontane Sprechweise. Sie entstehen beispielweise durch falsch angefangene, abgebrochene, neubegonnene oder wiederholte Wort– bzw. Satzfragmente. Störend auf den Sprachfluß wirken sich auch Stottern, Zögern, betonte, gedehnte oder falsch ausgesprochene Wörter aus.

Gegenstand dieser Arbeit sind Sprechpausen, d.h. Lücken, die durch Unterbrechungen im Sprachfluß zwischen den oben beschriebenen Wort– oder Satzfragmenten entstanden sind. Solche Sprechpausen können mit Geräuschen ausgefüllt oder nicht ausgefüllt sein. In dieser Arbeit interessieren besonders die mit Geräuschen ausgefüllten Sprechpausen. Unter den Begriff „Geräusche" fallen einerseits vom menschlichen Stimmorgan erzeugte nonverbale Produktionen (bspw. Lachen, Husten, Schmatzen, Atmen, Häsitationen ...), im weiteren *sprachliche Geräusche* genannt und andererseits nichtartikulatorische Geräusche (bspw. Papierrascheln, Tastaturklicken, Telefonklingeln ...), im folgenden als *nichtsprachliche Geräusche* bezeichnet.

Die oben erwähnten Merkmale spontaner Sprache vermindern die Leistung bestehender Spracherkennungs– und Übersetzungssysteme beträchtlich. Untersuchungen [1] haben gezeigt, daß die Erkennungsleistung eines Spracherkennungssystems auf spontan gesprochener Sprache signifikant schlechter wurde, als auf gelesener Sprache. Nach Ansicht der Autoren liegt die Hauptursache für die unterschiedlichen Fehlerraten zwischen spontaner und gelesener Sprache im Sprachfluß. Darunter fassen sie gefüllte Pausen, lange Pausen, Wortfragmente, betonte, gedehnte und falsch ausgesprochene Wörter sowie Abbrüche und Selbstkorrekturen. Innerhalb dieser Liste, sind gefüllte Pausen die zweithäufigste Ursache der höheren Fehlerraten bei spontaner Sprache. Insgesamt konnten je 20% der Fehler durch gefüllte Pausen und Geräusche sowie durch Grammatik- und Vokabularprobleme aufgeklärt werden. 30% der Fehler werden durch die Artikulation verursacht, die restlichen 30% der Fehler blieben ungeklärt. Aus dieser Untersuchung kann unmittelbar abgeleitet werden, daß die explizite Modellierung von mit Geräuschen gefüllten Pausen im Idealfall eine Verminderung der Fehlerrate um 20% erwarten läßt [1].

Ziel dieser Studienarbeit war es, die sprachlichen und nichtsprachlichen Geräuschen in Sprechpausen akustisch zu modellieren. Dadurch sollte die Erkennungsleistung des bestehenden Sprachübersetzungssystems JANUS für spontan gesprochene Sprache verbessert werden.

Das nächste Kapitel beschäftigt sich zur Erläuterung der Vorgehensweise bei der akustischen Modellierung zunächst mit einer Beschreibung des JANUS-Systems und hier insbesondere mit dem spracherkennenden Modul. Der nachfolgende Abschnitt beschreibt den Ansatz, der zur Lösung der gestellten Aufgabe gewählt wurde. Nach der Beschreibung des verwendeten Datenmaterials werden die Ergebnisse dargestellt und interpretiert.

2 Das Janus-System

Janus ist ein Sprach-zu-Sprach Übersetzungssystem und wird im Rahmen einer Kooperation zwischen der Universität Karlsruhe und der Carnegie-Mellon University, Pittsburgh entwickelt [2]. Das ganze System ist modular aufgebaut und besteht aus einem Spracherkenner, der sprecherunabhängig kontinuierlich gesprochene englische oder deutsche Sprache erkennt, einem Parser, der den vom Spracherkenner gelieferten Text analysiert und das Erkannte in eine künstliche Zwischensprache *Interlingua* übersetzt, sowie einer Einheit zur Sprachausgabe, die den in der Zielsprache generierten Satz in akustische Signale umwandelt und ausspricht.

Da sich die gestellte Aufgabe nur auf die akustische Modellierung bezieht, wird für die weiteren Ausführungen nur der spracherkennende Teil von Janus näher beschrieben.

2.1 Die Signalvorverarbeitung

Ein wichtiger Ausgangspunkt des Erkennens von Sprache ist die Klassifikation der Sprachdaten anhand geeigneter Merkmale. Die Natur (bspw. beim menschlichen Gehör) lehrt uns, daß der Energiegehalt logarithmisch aufgeteilter Frequenzbänder des Sprachsignals ein solches geeignetes Unterscheidungsmerkmal darstellt. Die benötigten Klassifikationsmerkmale werden beim Janus-System durch eine geeignete Vorverarbeitung aus dem aufgenommene Rohsignal extrahiert. Im Einzelnen besteht die Vorverarbeitung aus der Abtastung und Quantisierung des analogen Sprachsignals durch einen 14bit A/D-Wandler mit einer Abtastrate von 16 kHz. Anschließend werden diejenigen Sprachteile am Anfang und am Ende abgeschnitten, die unter einem definierten Aussteuerungspegel liegen (*clipping*). Danach wird eine Spektralanalyse durchgeführt, indem über die verbleibenden Sprachdaten äquidistant alle 5 ms ein Hamming-Fenster geschoben und eine Fouriertransformation berechnet wird. Aus den so entstehenden FFT-Koeffizienten werden Leistungskoeffizienten und daraus 16 melscale Koeffizienten ermittelt. Man erhält auf diese Weise für je 10 ms Sprache einen 16-dimensionalen Merkmalsvektor. Die Vektoren werden paarweise gemittelt und auf das Intervall $[0, 1]$ normiert. Der Erkenner bekommt dann als Eingabe für jeden gesprochenen Satz eine Datei alle auf diese Weise vorverarbeiteten Vektoren an einem Stück.

2.2 Die Analyse des Sprachsignals

Ziel der Spracherkennung ist es, aus dem gegebenen Sprachsignal möglichst zuverlässig und fehlerfrei Wortketten zu erkennen. Dies wird mit wachsendem Vokabular immer schwieriger, die Erkennung unsicherer. Aufgrund dieser Unsicherheit ist man in der automatischen Spracherkennung gezwungen, aus einer Vielzahl von möglichen Hypothesen diejenige auszuwählen, die unter vorgegebenen Randbedingungen am wahrscheinlichsten ist. Die statistische Entscheidungstheorie sagt darüber [3]:

> um die Wahrscheinlichkeit für einen Erkennungsfehler zu minimieren, muß man die Wortsequenzen $\hat{w_1} \ldots \hat{w_N}$ so wählen, daß $P(\hat{w_1} \ldots \hat{w_N} | x_1 \ldots x_T) = max_w P(w_1 \ldots w_N | x_1 \ldots x_T)$, d.h. es ist diejenige Sequenz von Wörtern $w_1 \ldots w_N$ mit unbekannter Länge N zu bestimmen, die die größte Wahrscheinlichkeit unter allen möglichen Wortsequenzen für die Beobachtungssequenz $x_1 \ldots x_T$ über der Zeit T hat.

Mit dem Satz von Bayes und der Tatsache, daß $P(x_1 \ldots x_T)$ im Vergleich der Hypothesen untereinander konstant ist, ergibt sich daraus:

$$max P(w_1 \ldots w_N | x_1 \ldots x_T) = max[P(w_1 \ldots w_N) \cdot P(x_1 \ldots x_T | w_1 \ldots w_N)] \quad (1)$$

Der erste Ausdruck $P(w_1 \ldots w_N)$ der Gleichung (1) beschreibt die sprachliche Modellierung (*Language modelling*). Hierbei ist $P(w_i)$ die a-priori Wahrscheinlichkeiten, sie gibt die Auftrittswahrscheinlichkeit für ein einzelnes Wort an. Analog dazu bezeichnet $P(w_1 \ldots w_N)$ die Auftrittswahrscheinlichkeit für eine bestimmte Wortfolge. Hierdurch werden syntaktische und teilweise semantische Randbedingungen festgelegt. Bei dem hier verwendeten Erkenner wurde eine Wortpaar–Grammatik benutzt, die auf den Sätzen der Testmenge erstellt wurde.

Der zweite Ausdruck $P(x_1 \ldots x_T | w_1 \ldots w_N)$ der Gleichung (1) beschreibt die akustisch–phonetische Modellierung, d.h. die Wahrscheinlichkeit dafür, daß die akustischen Vektoren $x_1 \ldots x_T$ beobachtet werden, wenn die Wörter $w_1 \ldots w_N$ gesprochen wurden. Wobei die Beobachtungssequenz $x_1 \ldots x_T$ den akustischen Vektoren entspricht, die bei der oben beschriebenen Vorverarbeitung entstanden sind. Die Wahrscheinlichkeiten werden während des Trainings geschätzt. Auf welche Weise dies bei dem hier verwendeten JANUS-System geschieht, wird im folgenden Abschnitt erläutert.

Für Spracherkennungssysteme mit großen Vokabularen werden typischerweise Einheiten verwendet, die kleiner sind als Wörter. JANUS benutzt als Basiseinheit für den Erkenner Phoneme. Die Wortmodelle erhält man dann durch Konkatenation der Phoneme. Zu diesem Zweck existiert ein Wörterbuch, das eine phonetische Umschrift aller Wörter des Vokabulars enthält und bei dem verwendeten System auf 44 Phonemen für die englischen Sprachlaute basiert.

Schließlich muß eine Entscheidung darüber gefällt werden, was insgesamt im ganzen Satz gesprochen wurde. Man benötigt hierzu eine Suchstrategie, die alle verfügbaren Wissensquellen zu einem optimalen Ergebnis kombiniert. Beachtet werden müssen also gleichzeitig das Sprachmodell, das Wissen über die Konkatenation der Phoneme zu ganzen Wörtern und das akustisch-phonetische Modell. Zur Suche der besten Hypothesen verwendet man den *Dynamic Time Warping (DTW)* Algorithmus, auch *dynamisches Programmieren* genannt. Dabei wird zum gesprochenen Eingabesatz diejenige Satzhypothese gesucht, die insgesamt unter allen festgelegten Randbedingungen die beste Bewertung erreicht. Im JANUS-System wird dazu der sogenannte *Word-dependent N-Best* Algorithmus verwendet. Weitere Ausführungen über die verwendeten Prinzipien würden den Rahmen der vorliegenden Arbeit sprengen, der interessierte Leser wird an dieser Stelle auf die entsprechende Literatur verwiesen [4].

2.3 Akustisch–phonetische Modellierung

Die akustisch–phonetische Modellierung in JANUS, die für die vorliegende Studienarbeit verwendet wurde, basiert auf einem hybriden Ansatz, einer Kombination eines neuronalen Ansatzes, dem *Learning Vector Quantisation (LVQ)* und einem probabilistischen Ansatz, dem *Hidden Markov Model (HMM)*. Dieser Ansatz wurde in [5] vorgestellt und soll die Vorteile des LVQ, optimale Klassifikation, mit den Vorteilen des HMM, gute Zeitmodellierung von Sprache, verbinden.

Zunächst werden die Merkmalsvektoren der Trainingsbeispiele mit dem *k-means*-Algorithmus geclustert und anhand der entstandenen Vektoren ein sogenanntes Codebuch mit einer festen Anzahl Referenzvektoren erstellt. Dieses initiale Codebuch ist Ausgangspunkt für den *LVQ-2*-Algorithmus, der die Referenzvektoren im Training so verschiebt, daß jeder Merkmalsvektor einen Referenzvektor der richtigen Klasse als nächsten Nachbarn bekommt. Der *LVQ-2*-Algorithmus minimiert somit die Anzahl der Fehlklassifikationen.

Die beim JANUS-System verwendeten Phonemmodelle werden als HMMs interpretiert, wodurch die zeitliche Struktur der Sprache modelliert wird. Da sich Anfangs-, Mittel- und Endabschnitt eines Phonems akustisch stark voneinander unterscheiden können, wird das Phonemmodell in 3 Segmente unterteilt. Geht man von einer durchschnittlichen Länge eines Phonems von 60 ms aus und berücksichtigt, daß ein Merkmalsvektor 10 ms lang ist, so ergeben sich für das HMM-Phonemmodell 6 Zustände. Um die Zahl der zu schätzenden HMM-Parameter gering zu halten, wird je zwei Zuständen eine gemeinsame Emissionswahrscheinlichkeitsverteilung zugeordnet.

Im Training wird dann das erzeugte Codebuch verwendet, um anhand des *forward-backward*-Algorithmus für HMMs die Trainingsbeispiele in Sequenzen von Codebuch-Indizes zu zerlegen, anhand derer dann die HMM-Parameter geschätzt werden.

Beim Test eines unbekannten Satzes erhält der Erkenner als Eingabe die FFT-Merkmalsvektoren des gesprochenen Satzes und liefert für alle Phoneme je eine Bewertung für jedes der drei Segmente, mit denen dann die DP-Matrix initialisiert wird.

3 Der Lösungsansatz

Wilpon et al. [6] setzten erstmals in einem HMM-basierten Keyword-spotter für nicht bekannte und nichtsprachliche akustische Ereignisse spezielle HMM-Modelle ein. Die Autoren zeigten, daß durch die separate Modellierung der vier Geräusche *Hintergrund-Geräusch*, *Lippenschmatzen und Telefonleitungs-Knacklaute*, *Atemgeräusche* und *Hintergrundkonversation* eine Verbesserung der Erkennungsleistung ihres Systems erzielt werden konnte.

W. Ward [7],[8] stellte für das ebenfalls HMM-basierte System PHOENIX einen Ansatz vor, bei dem zusätzlich zu den normalen Phonemmodellen für Wörter 14 spezielle Modelle hinzugefügt wurden, die ausschließlich Geräusche repräsentieren. Diese Geräuschmodelle wurden genauso gehandhabt wie Wortmodelle, nur wurden sie nicht mit Wörtern sondern mit Geräuschen trainiert. Außerdem wurden sie durch kein Sprachmodell eingeschränkt, sondern durften nach jedem Wort und nach Geräuschen auftreten. Ihr Einsatz brachte bei geräuschbehafteter Spracheingabe eine prozentuale Verminderung des Worterkennungsfehlers um 55 % gegenüber dem System ohne Geräuschmodelle.

Dieser Ansatz wurde auch in der vorliegenden Studienarbeit verfolgt. Entsprechend der Architektur des Janus-Systems wird ein Geräusch durch ein eigenes Phonemmodell modelliert, d.h. jedes Geräusch wird im Wörterbuch durch ein einziges Phonem beschrieben. Allerdings werden - im Gegensatz zum Vorgehen von Ward - die Geräuschmodelle explizit in die verwendete Wortpaar-Grammatik eingebaut.

Anhand der erzeugten Geräuschmodelle sollte die Frage beantwortet werden, ob eine Clusterung von Geräuschen zu gemeinsamen Modellen eine Steigerung der Erkennungsleistung erbringt. Diese ist zu erwarten, weil durch eine Clusterung das Trainingsmaterial auf weniger Modelle konzentriert und dadurch besser ausgenutzt wird. Besonders bei sehr wenig Trainingsmaterial kommt diesem Umstand eine große Bedeutung zu. Ein weiterer Vorteil, der bei einer minimalen Anzahl von Geräuschmodellen entsteht, ist der geringere Rechenaufwand und ein kleinerer Speicherplatzbedarf.

Zur Lösung der Aufgabe wurde im Einzelnen folgendermaßen vorgegangen:

1. Transkribieren aller Äußerungen inklusive Geräusche
2. Bilden von Geräuschklassen
3. Eintrag eines speziellen Phonems pro Geräuschklasse in die Phonemliste
4. Aufnehmen aller Geräuschklassen in das Wörterbuch
5. Integrieren der Geräuschklassen in die Grammatik
6. Trainieren auf den mit Geräuschklassen transkribierten Äußerungen
7. Clustern der Geräuschklassen
8. Wiederholen der Schritte 3,4,5,6 für die verschiedenen Varianten der Geräuschcluster
9. Testen der verschiedenen Geräuschclustervarianten

Bei dem zu Verfügung stehenden Datenmaterial lagen die Transkriptionen inklusive der Geräusche vor, so daß Schritt 1 der Liste bereits bearbeitet war. Kapitel 4 beschreibt das verwendete Material und die Vorgehensweise zu den Schritten 2, 3 und 4. Kapitel 5 enthält die Angaben zu Schritt 6, 7 und 8. Die Ergebnisse der Schritte 1 bis 9 sind im Kapitel 6 dargestellt.

4 Zugrundeliegendes Datenmaterial

Ursprünglich sollte die Analyse anhand des **A**ir **T**ravel **I**nformation **S**ervice Datensatzes durchgeführt werden. Dieser Datensatz enthält spontan gesprochene, englische Anfragen an ein Datenbanksystem, das Auskunft über Fluginformationen gibt. ATIS wird von der DARPA unterstützt und von den DARPA-geförderten Einrichtungen zur Entwicklung und Evaluation ihrer Systeme benutzt [9]. Der zur Verfügung stehende Datensatz umfaßt mehr als 10000 Sätze. Eine eingehende Analyse des Datenmaterials ergab jedoch, daß insgesamt zu wenig Geräuschmaterial zur Verfügung stand, um eine zuverläßige Ermittlung aller notwendigen Parameter zu gewährleisten. Die akustische Modellierung der Geräusche wurde daher anhand des gerade entstehenden **E**nglish **S**pontaneous **S**cheduling **T**ask durchgeführt. Beim ESST handelt sich um spontansprachliche Dialoge zwischen zwei Personen. Die Geprächspartner haben die Aufgabe, einen gemeinsamen Termin für ein (fiktives) Treffen zu arrangieren und erhalten zu diesem Zweck einen Kalender, wodurch das Szenario eingegrenzt wird. Diese Daten werden derzeit in Kooperation von Universität Karlsruhe und der Carnegie Mellon University gesammelt [10].

4.1 Vorhandene Geräusche

Zur Verfügung standen 43 Trainings- und 20 Testdialoge, die insgesamt aus 447 Trainings- bzw. 190 Testäußerungen bestehen. Das Wörterbuch umfaßt 865 Einträge. Die Äußerungen wurden auf die Zahl der Vorkommen von sprachlichen und nichtsprachlichen Geräuschen hin untersucht. Tabelle 1 faßt die Häufigkeiten für die sprachlichen Geräusche, Tabelle 2 die für die nichtsprachlichen Geräusche zusammen.

Zusätzlich zu den sprachlichen und nichtsprachlichen Geräuschen wurde noch ein Geräusch `+MUELL+` eingeführt, auf das Wortfragmente abgebildet wurden, die durch Verstümmelungen (Wiederholung, Neuanfang oder Abbruch) entstanden sind und so stark von der richtigen Aussprache abwichen, daß sie nicht mehr rekonstruiert werden konnten. Solche Phänomene traten im Training 78 und im Test 26 mal auf. Verstümmelte Wörter wurden als korrekt ausgesprochen in die Analyse aufgenommen, sofern das Transkribierte nicht mehr als 2 Zeichen von der richtigen Schreibweise abwich.

4.2 Bilden von Geräuschklassen

Um ein Phonem modellieren zu können, muß eine Mindestanzahl von Vorkommen gewährleistet sein, damit genügend Trainings- aber auch Testmaterial vorhanden ist. Daher sollten nur solche Geräusche modelliert werden, die häufig genug im Training auftraten. Zu diesem Zweck wurden Geräuschklassen gebildet. Alle Geräusche, die mindestens 50 mal in der Trainingsmenge auftraten, bilden eine Geräuschklasse für sich (`+AH+` wurde als Grenzfall noch als Einzelmodell aufgenommen). Die restlichen sprachlichen

Bezeichnung	Geräusch	# Training	# Test
+AH+	Füllwort, das wie ah klingt	48	24
+EH+	Füllwort, das wie eh klingt	5	0
+GLOTTAL+	Verschlußlaut	33	7
+GULP+	Schlucken	0	1
+H#+	Atemgeräusch	1069	364
+HM+	Füllwort, das wie hm klingt	13	1
+HUH+	Füllwort, das wie huh klingt	3	1
+LG+	Lachen	26	22
+LS+	Lippenschmatzer	458	162
+MM+	Füllwort, das wie mm klingt	4	0
+OH+	Füllwort, das wie oh klingt	14	3
+OO+	Füllwort, das wie oo klingt	4	1
+UH+	Füllwort, das wie uh klingt	280	96
+UM+	Füllwort, das wie um klingt	298	78
+YAH+	Füllwort, das wie yah klingt	0	1

Tabelle 1: sprachliche Geräusche in 447 Trainings- und 190 Testäußerungen

Geräusch	#Training	#Test
Binder_Closing	1	0
Chair_Squeak	3	0
Copier_Lid_Falling	4	0
Copier_Noise	9	3
Door_Slam	2	1
Drawer_Slam	1	0
Finger_Hitting_Headset	2	1
Foot_Kicking_Trashcan	1	0
Headset_Adjustment	4	0
Headset_Hitting_Table	1	0
Key_Click	469	204
Microphone_Adjustment_Noise	2	0
Microphone_Noise	21	12
Mouse_Click	1	0
Mug_Hits_Table	1	0
Object_Hitting_Table	1	1
Paper_Rustle	61	24
Paper_Sliding	2	0
Pen_Tap	12	1
Phone_Ring	1	1
Terminal_Beep	4	0
Twang	1	0
Writing_Noise	1	0

Tabelle 2: nichtsprachliche Geräusche in 447 Trainings- und 190 Testäußerungen

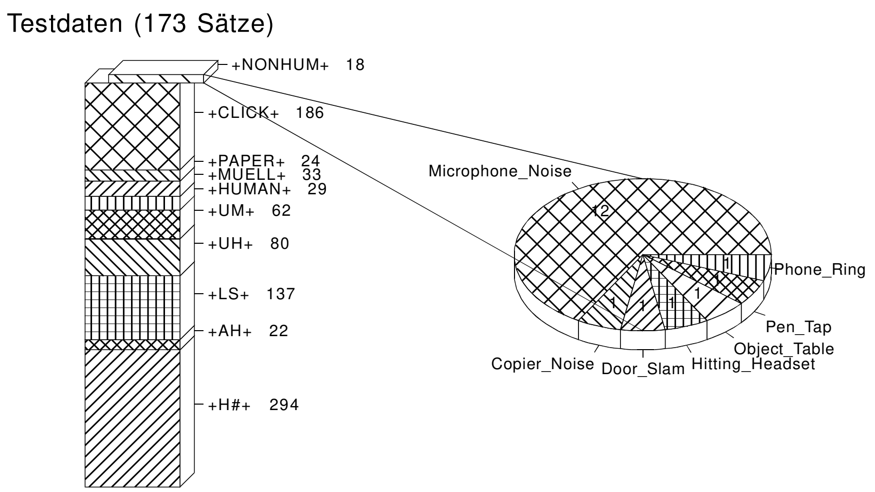

Abbildung 1: Auftrittshäufigkeiten der 10 Geräuschklassen in 173 Testäußerungen

Geräusche wurden zu einer gemeinsamen Klasse **+HUMAN+**, die nichtsprachlichen zu einer gemeinsamen Geräuschklasse **+NONHUM+** zusammengefaßt. Insgesamt ergaben sich inklusive des Modells **+MUELL+** 10 Geräuschklassen.

Aus technischen Gründen konnten nicht alle Sätze in die Verarbeitung aufgenommen werden, so daß letztendlich noch 387 Trainings- und 173 Testäußerungen in der Analyse verblieben. Das entspricht insgesamt 62 Minuten Sprache für das Training und 27 Minuten Sprache für das Testen. Die Geräusche bilden einen Anteil von 20 % aller in der Testmenge geäußerten Wörter. Die Grafiken zeigen die Auftrittshäufigkeiten der Geräuschklassen im Test und im Training.

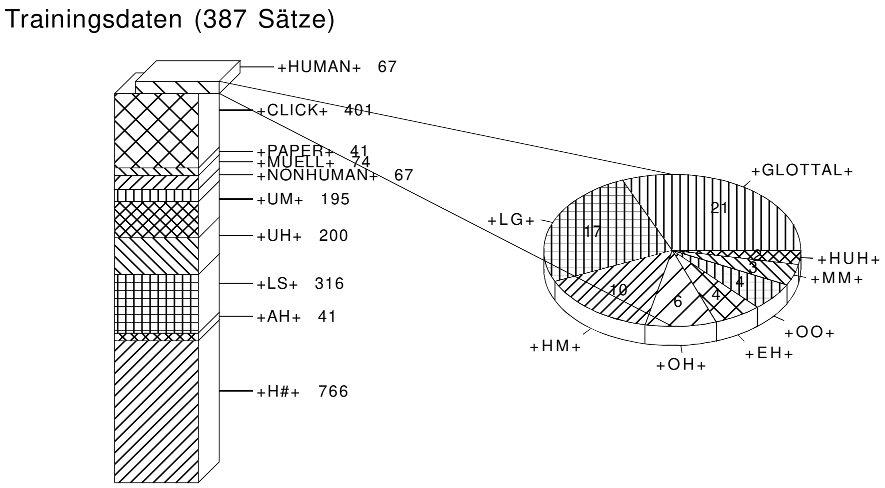

Abbildung 2: Auftrittshäufigkeiten der 10 Geräuschklassen in 387 Trainingsäußerungen

5 Entwicklung der Geräuschmodelle

Die Entwicklung der Modelle sollte nun, wie in Abschnitt 3 beschrieben, anhand des dargestellten Datenmaterials durchgeführt werden. Die hierfür notwendigen Schritte sind im einzelnen:

1. Erstellen eines Codebuches für ein globales Geräuschmodell als Ausgangsbasis für alle Geräuschklassenmodelle
2. Trainieren aller Verteilungen der 10 Geräuschklassen
3. Clustern der Geräuschklassen aufgrund der Verteilungen
4. Trainieren und Testen der verschiedenen Varianten von Geräuschclustern
5. Bestimmung der optimalen Clustervariante

5.1 Erstellen eines Codebuches

Ziel war zunächst die Erzeugung eines Phonemmodells für alle in Abschnitt 4 definierten Geräusche. Zu diesem Zweck wurde zunächst ein Codebuch für ein allgemeines Geräuschmodell entwickelt. Dazu wurde ein Phonem in die Phonemliste aufgenommen, dem alle Geräusche zugewiesen wurden. Die durch den *Viterbi*-Algorithmus den Geräuschen zugewiesenen Merkmalsvektoren wurden mittels des *k-means Clustering*-Algorithmus auf die für das Codebuch übliche Größe von 50 Referenzvektoren geclustert. Das entstandene Codebuch enthält somit Merkmalsvektoren für alle vorkommenden sprachlichen und nichtsprachlichen Geräusche und bildet den Ausgangspunkt für das weitere Vorgehen.

5.2 Trainieren der Verteilungen über alle Modelle

Im zweiten Schritt wurden für jede der in Abschnitt 4 beschriebenen Geräuschklassen ein eigenes Phonem in die Phonemliste eingetragen. Jedes Geräusch einer Klasse besteht somit aus einem einzelnen Phonem. Dabei verwendet jedes Geräusch das unter 5.1 erstellte Codebuch, lediglich die Verteilungen auf dem Codebuch wurden individuell trainiert. Der Erkenner wurde einige Iterationen mit diesen individuellen Geräuschmodellen trainiert. Ausgangsbasis war dabei ein Erkenner, bei dem die übrigen Phoneme für Wörter bereits gut trainiert waren.

5.3 Clustern der Geräuschklassen

Das größte Problem bei der Arbeit mit Geräuschmodellen ist derzeit der Mangel an Trainingsmaterial. Man strebt daher bei vorgegebenen Trainingsmenge eine Ausgewogenheit zwischen maximaler Sensitivität der Modelle auf der einen Seite und optimaler Trainierbarkeit auf der anderen Seite an. Die Idee war, die 10 Geräuschklassen nach Ähnlichkeit zu clustern um auf diese Weise das Trainingsmaterial auf weniger Modelle zu konzentrieren. Es wurde erwartet, daß durch das bessere Training der Modelle die Einbußen in der Erkennungsleistung durch weniger sensitive Modelle mehr als aufgewogen werden. Eine Veringerung der Zahl der Modelle führt darüberhinaus zu zweckmäßigeren Systemen im Sinne des Speicherbedarfs und Rechenzeitaufwandes.

Zum Mischen der Modelle wurde ein Algorithmus verwendet, der bei K.F. Lee [11] zum Triphone Clustering vorgeschlagen wurde.

1. Bilde für jede Geräuschklasse ein eigenes Cluster
2. Mische das ähnlichste Paar von Clustern zusammen
3. Bewege jedes Element eines Clusterpaares von einem Cluster zum anderen und führe die Bewegung durch, sofern dabei eine Verbesserung der Konfiguration entsteht. Dieser Schritt wird solange wiederholt, bis keine solche Bewegung mehr übrig ist
4. Wiederhole ab Schritt 2 bis ein Konvergenzkriterium erfüllt ist

Als Maß für die Ähnlichkeit zweier Geräuschklassen wird eine informationstheoretisches Maß verwendet: die Information, die verloren geht, wenn man zwei Modelle mischt, gemessen an der Differenz der Entropie $H_{x,d}$ zwischen den ursprünglichen Modellen a, b und dem gemischten Modell m. Der Informationsverlust wird gewichtet mit der Auftrittshäufigkeit $N_{x,d}$ der Modelle. Die Distanz berechnet sich somit zu:

$$L(a,b) = \sum_d [N_{m,d}H_{m,d} - (N_{a,d}H_{a,d} + N_{b,d}H_{b,d})] \tag{2}$$

wobei $H_{x,d} = -\sum_i P_{x,d}(i) \cdot log(P_{x,d}(i))$ die Entropie der Ausgabewahrscheinlichkeit für die Verteilung d ist und $P_{x,d}(i) = N_{x,d}(i)/N_{x,d}$. $N_{x,d} = \sum_i N_{x,d}(i)$ und $N_{x,d}(i)$ ist die Häufigkeit des Auftretens von Codebuchvektor i in der Verteilung d des Kontextes x, wie sie *im forward-backward* Algorithmus gezählt wird. Unter d versteht man hier die 6 Verteilungen, je drei über die Phonemsegmente für die zwei LVQ-Netze pro Phonem, x steht für die Geräuschklassen a, b, m.

Die Grafik zeigt die Ergebnisse des Clusteralgorithmus. Man sieht, daß durch das gewichtete Maß diejenigen Geräuschmodelle bevorzugt zusammengemischt werden, die seltener auftreten. Auf diese Weise erreicht man, daß diese Modelle mehr Trainingsmaterial erhalten.

5.4 Trainieren und Testen der Clustervarianten

Für jede der im vorigen Abschnitt ermittelten Clustervarianten wurde nun jeweils eine Phonemliste, ein Wörterbuch, eine Grammatik und Transkriptionen erstellt und darauf jeweils 20 Iterationen trainiert. Dabei wurde im Abstand von zwei bzw. einer Iteration getestet.

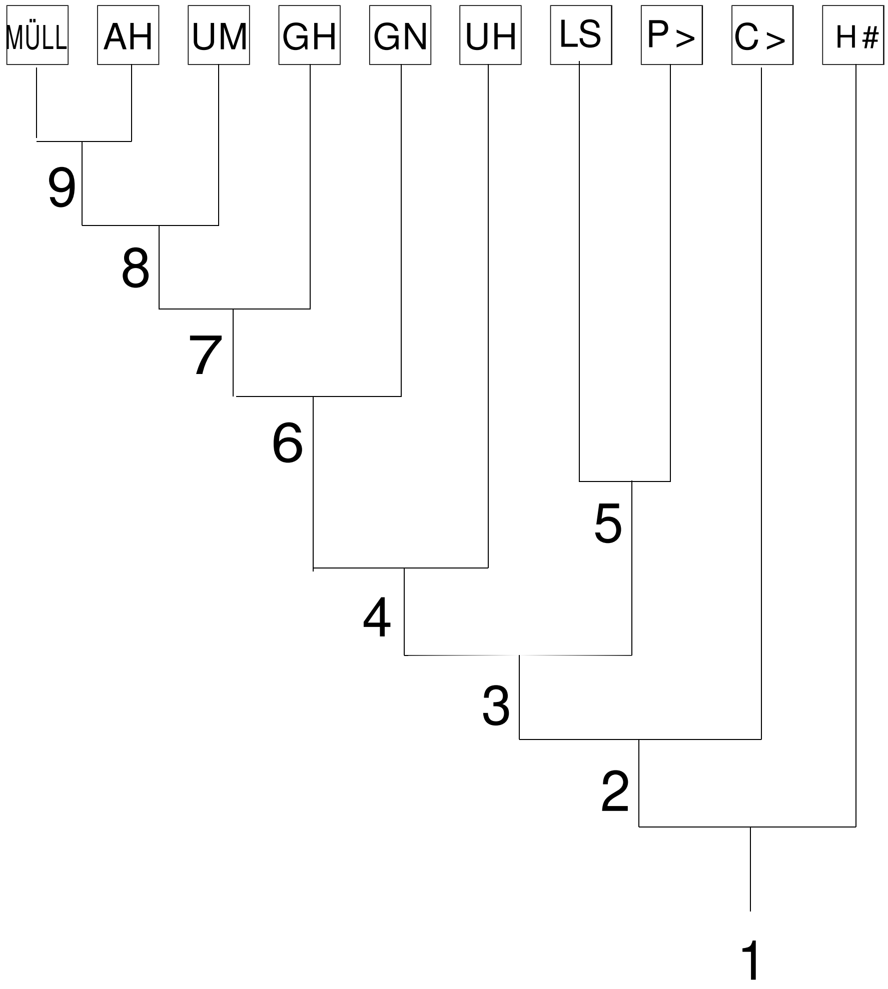

Abbildung 3: Clustervarianten der Geräuschklassen

6 Ergebnisse

Im vorliegenden Abschnitt soll nun beschrieben werden, welchen Einfluß die unterschiedlichen Clustervarianten der Geräusche sowie deren Training auf die Erkennungsleistung des JANUS-Systems ausüben.

6.1 Maßstab für die Erkennungsleistung

Wie in der Spracherkennung allgemein üblich, wurde als Maß für die Erkennungsleistung des akustischen Teils der sogenannte Wort-Erkennungsfehler (*Word Error* WE) herangezogen. Zur Berechnung wird dabei die Transkription des vom Erkenner hypothetisierten Satzes mit der Transkription des tatsächlich gesagten Referenzsatzes Wort für Wort verglichen. Der Wort-Erkennungsfehler besteht aus der Summe der Wortverwechslungen *(substitutions)*, fälschlichen Worteinfügungen *(insertions)* und fälschlichen Wortauslassungen *(deletions)* bezogen auf alle Referenzwörter in Prozent. Die Wort-Erkennungsleistung (*Word Accuracy* WA) berechnet sich daraus zu $WA = 100\% - WE$.

Ziel der Geräuschmodellierung ist es, die vorhandenen Geräusche zu erkennen, damit sie anschließend eliminiert werden können. Da die Geräusche keine semantische Bedeutung tragen, ist es nicht sinnvoll, sie zu Weiterverarbeitung in den Sätzen zu belassen. Außerdem sollten Geräusche bei einem Übersetzungsssystem als nicht übersetzungsrelevante Teile angesehen werden. Aus diesem Grund wurde die Erkennungsleistung nicht auf der Basis der tatsächlich erkannten Satzhypothesen berechnet, sondern auf den Sätzen, die vorher von allen Geräuschen befreit wurden. Entsprechend wurden auch alle Geräuschvorkommen aus den Transkriptionen der Referenzsätze entfernt.

Beispiel

vor der Elimierung aller Geräusche:

Ref	+LS+	okay	+H#+	so	i'll		hear	you	+MUELL+	monday	thanks	+CLICK+
Hyp	+LS+	okay		so	i'll	+AH+	see	you	on	monday	+UH+	+PAPER+

nach der Elimierung aller Geraäusche:

Ref		okay		so	i'll		hear	you		monday	thanks	
Hyp		okay		so	i'll		see	you	on	monday		
			(A)			**(B)**	**(C)**		**(D)**		**(E)**	**(F)**

Wortverwechslungen werden daher nur gezählt, wenn ein Wort des hypothetisierten Satzes mit einem Wort des Referenzsatzes verwechselt wurde (Fall C im obigen Beispiel). Wird dagegen ein Referenzwort fälschlicherweise als Geräusch hypothetisiert, so wird dieser Erkennungsfehler nicht als Wortverwechslung sondern als Wortauslassungsfehler gewertet, weil durch die vorherige Eliminierung des fehlerhaft eingesetzten Geräusches im Hypothesensatz anstelle einer Hypothese nun eine Lücke klafft (Fall E).

Umgekehrt entsteht ein Worteinfügungsfehler, wenn ein Geräusch im Referenzsatz fälschlicherweise als Wort im Hypothesensatz erkannt wird. Nach Eliminierung des Referenzgeräusches erscheint das hypothetisierte Wort, welches nicht von der Eliminierung betroffen ist, als Einschub (Fall D im Beispiel). Verwechslungen der Geräusche untereinander (Fall F) werden nicht als Fehler gewertet. Dies ist insofern nicht problematisch, als das Ziel der Geräuschmodellierung nicht darin besteht, Geräusche voneinander abzugrenzen, sondern vielmehr darin, sie von semantisch bedeutungsvollen akustischen Ereignissen zu unterscheiden. Weiterhin werden fehlerhafte Einfügungen (Fall B) und fehlerhafte Auslassungen (Fall A) von Geräuschen nicht beachtet. Sprunghaftes Ansteigen dieser beiden Fehlertypen führt so zwar nicht zu einer Verringerung der Worterkennungsleistung, ist aber als Indiz für Schwächen in der Akustik der Geräuschmodelle zu werten und daher zu vermeiden.

Wenn also im folgenden von Erkennungsleistung die Rede ist, ist damit immer der Vergleich zwischen geräuschbereinigten Hypothesensätzen und geräuschbereinigten Referenzsätzen gemeint.

6.2 Ergebnisse der einzelnen Geräuschcluster

Die Abbildung 5 und die zugehörige Tabelle 3 stellen die Worterkennungsleistung auf der Testmenge für jede der 10 Geräuschclustervarianten dar. Zur grafischen Veranschaulichung wurden die Erkennungsleistungen geglättet

$$WA(i) = [WA(i-1) + WA(i) + WA(i+1)]/3 \qquad mit \quad i = 37 \ldots 57 \quad (3)$$

Die Angaben in der Tabelle enthalten das arithmetrische Mittel der Worterkennungsraten über die gekennzeichneten Intervalle.

Iteration	36	37 - 43	44 - 50	51 - 57	58
1 CL	42.7	44.6	45.9	46.4	47.2
2 CL	42.4	45.4	46.0	45.1	43.7
3 CL	42.1	45.3	47.0	48.6	48.7
4 CL	42.2	45.6	47.1	45.0	44.8
5 CL	41.9	45.5	48.1	49.1	49.1
6 CL	41.8	45.2	46.4	50.4	51.1
7 CL	41.8	45.2	44.7	44.0	43.2
8 CL	41.6	44.9	44.6	46.1	45.4
9 CL	41.5	44.3	47.4	48.6	49.6
10 CL	41.4	44.2	47.1	49.0	49.8

Tabelle 3: ungeglättete Worterkennungsleistung aller Clustervarianten

Abbildung und Tabelle zeigen die Entwicklung der Erkennungsleistung in Abhängigkeit der Trainingsiterationen über alle 20 Iterationen hinweg, wobei jede 2. Iterationen und ab Iteration 52 jede Iteration ein Testpunkt liegt. Dargestellt sind also 10 Varianten an 15 Meßzeitpunkten, d.h. 150 Meßwerte. Es muß darauf hingewiesen werden, daß es sich bei der Grafik um einen kleinen Ausschnitt der Y-Achse handelt, so daß die Verbesserungsraten insgesamt etwas dramatischer wirken, als sie tatsächlich sind. Trotzdem wurde diese Darstellungsweise gewählt, um die Unterschiedlichkeit der Varianten aufzuzeigen.

Cluster 1 beschreibt die Entwicklung für ein allen Geräuschen gemeinsames Modell. Die gesamte Trainingsmenge wird somit auf ein einziges Geräuschmodell konzentriert, was maximales Training bedingt. Man sieht, daß die Erkennungsleistung im mittleren Bereich bezogen auf die übrigen Varianten liegt. Cluster 10 beschreibt das andere Extrem. Hier wird jede der 10 Geräuschklassen getrennt modelliert, wodurch maximale Sensitivität erreicht werden könnte, vorausgesetzt es steht genügend Trainingsmaterial zu Verfügung. Durch Aufteilung des Material auf 10 Modelle verfügt man im Schnitt nur noch über 10% des ursprünglichen Materials. Cluster 2 bis Cluster 9 sind die Zwischenlösungen der beiden Extreme.

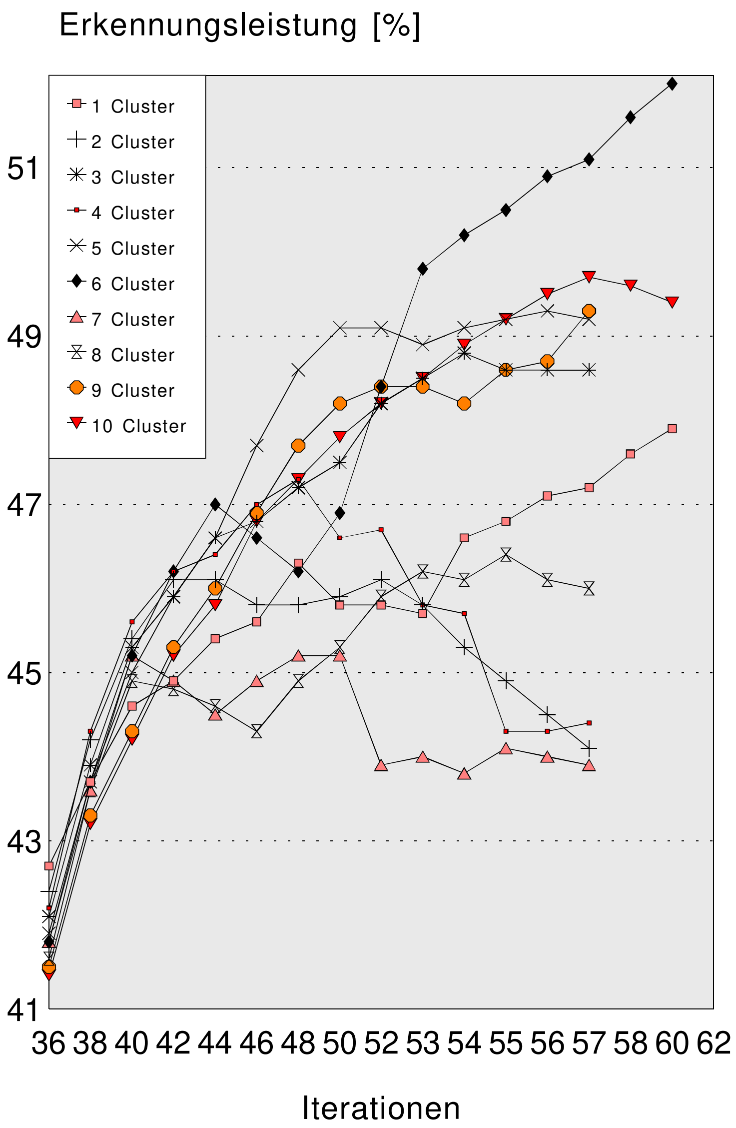

Abbildung 4: geglättete Worterkennungsleistung aller Clustervarianten

Wortfehlertyp	ohne Geräusche	mit Geräuschen	Differenz
Substitutions	29.1 % (1096)	32.2 % (1522)	3.1 % (426)
Deletions	8.0 % (301)	9.0 % (426)	1.0 % (125)
Insertions	10.5% (395)	12.9% (610)	2.4 % (215)
WA	52.5 %	45.9 %	6.6 %

Tabelle 4: Analyse der Wortfehlerraten

Wie aus der Grafik zu entnehmen ist, erzielt man mit der Variante, die aus 6 Clustern besteht, das mit Abstand beste Ergebnis. Es handelt sich dabei um die Variante, bei der Atemgeräusche `+H#+`, Papierrascheln `+PAPER+`, Tastaturklicken `+CLICK+`, Lippenschmatzen `+LS+` und das Füllwort `+UH+` getrennt modelliert, und alle restlichen Geräusche in eine gemeinsame Klasse zusammengefügt wurden. Offensichtlich ist diese Variante der in Abschnitt 5.3 erwähnte Kompromiß zwischen optimaler Sensitivität der Modelle und noch ausreichender Trainierbarkeit der Modelle. Zu klären bleibt, warum die Variante 6 besser ist als alle anderen Clustervarianten. Ein Vergleich der verschiedenen Varianten zeigte, daß die Lösung mit 6 Clustern vor allem bezüglich der Worteinfügungsfehler besser abschneidet als alle übrigen Varianten. Auf welchen Umstand dies zurückzuführen ist, soll der folgende Abschnitt beleuchten.

6.3 Analyse der Fehlerquellen

Für eine differenzierte Beurteilung der Folgen von den Geräuschmodelle auf die Erkennungsleistungen können die Auswirkungen, die sich beim Entfernen der Geräusche vor Berechnung der Erkennungsleistung ergeben, genutzt werden. Dazu wurde im ersten Test das beste Hypothesenfile (WI 6 Cluster 60 Iterationen) geräuschbereinigt mit dem geräuschbereinigten Referenzsätzen verglichen („ohne Geräusche“), im 2. Test wurden die Geräusche vor dem Vergleich nicht entfernt („mit Geräuschen“). Die Tabelle 5 stellt die Ergebnisse gegenüber.

Die Differenz bezüglich des Wortverwechslungsfehlers kommt nur durch Verwechslungen von Geräuschen untereinander zustande (vgl. Abschnitt 6.1). Insgesamt werden 426 mal zwei Geräusche miteinander verwechselt.

Da nur insgesamt 959 Geräusche in der Testmenge auftauchen, heißt dies, daß in 44 % der Fälle ein falsches Geräusch hypothetisiert wird. Dies deutet auf eine noch unzureichende Qualität der Geräuschmodelle hin, was als eine Folge des Mangels an Trainingsmaterial gewertet werden muß.

Die Differenz im Wortauslassungsfehler kommt zustande, wenn ein Wort aus dem Referenzsatz fälschlicherweise als Geräusch hypothetisiert wird. Dies passiert insgesamt in 125 Fällen von 3772 Referenzworten, d.h. in 3.3% der Fälle. In [1] wurde davor gewarnt, daß Modellen für gefüllte Pausen in vielen Fällen kurze Funktionswörter ersetzen würden. Diese Annahme kann aufgrund der Beobachtungen verworfen werden.

Die Differenz beim Worteinfügungsfehler kommt dadurch zustande, daß ein Geräusch im Referenzsatz fälschlicherweise als Wort hypothetisiert wird. Dies passiert in 215 Fällen. Bei einem Vorkommen von 959 Geräuschen in den Testsätzen entspricht das 22.5% und geht hier hauptsächlich zu Lasten kurzer Wörter (`the, to, two, but, time, for, i, you` ...).

Insgesamt kann festgestellt werden, daß die Hauptquelle der Unterschiede zwischen dem Test "mit Geräuschen" und "ohne Geräusche" in den Wortverwechslungen der Geräusche untereinander zu finden ist. Diese kann aber nur behoben werden, wenn genügend Trainingsmaterial zur Verfügung steht. Dazu besteht jedoch begründete Hoffnung, da die hier verwendete Datenbasis ESST derzeit an der Universität Karlsruhe ständig erweitert wird.

6.4 Signifikanz der Ergebnisse

Abschließend sollte untersucht werden, inwieweit die beobachteten Leistungsunterschiede zwischen den verschiedenen Clustervarianten statistisch bedeutsam sind. Die übliche Methode wäre ein Kreuzvalidierungstest. Dieser konnte jedoch wegen der geringen Auftrittshäufigkeiten von Geräuschen nicht durchgeführt werden. Es wurde daher ein neuer Weg beschritten, bei dem für jede Clustervariante die über die Testsätze gemittelte Anzahl der Wortfehler berechnet und anhand eines T-Tests die Signifikanz der beobachteten Mittelwertsunterschiede ermittelt wurde. Der T-Test berechnet die Wahrscheinlichkeit, daß der Mittelwert einer Differenz von 0 verschieden ist, wobei von normalverteilten bzw. bei kleinen Stichprobenumfänge t-verteilten Diffe-

Cluster–Variante	Iter 36 mean	Iter 36 p	Iter 50 mean	Iter 50 p	Iter 56 mean	Iter 56 p
1 Cluster	12,48	**(+)	11,61	ns	11,52	**(-)
2 Cluster	12,55	*(+)	11,68	ns	12,15	**(-)
3 Cluster	12,63	ns	11,45	*(+)	11,25	**(-)
4 Cluster	12,61	ns	11,46	*(+)	12,23	**(-)
5 Cluster	12,70	ns	11,02	**(+)	11,06	*(-)
6 Cluster	12,68		11,80		10,72	
7 Cluster	12,68	ns	12,20	*(-)	12,09	**(-)
8 Cluster	12,74	ns	11,91	ns	11,52	**(-)
9 Cluster	12,76	ns	11,12	**(+)	11,11	**(-)
10 Cluster	12,78	ns	11,39	*(+)	11,00	*(-)

Tabelle 5: statistische Bedeutsamkeit der Ergebnisse

renzen ausgegangen wird. Es wird dabei berücksichtigt, daß die Varianz der einen Meßwertreihe die der anderen mitbeeinflußt (abhängige Messungen).

Die Tabelle zeigt die mittlere Anzahl der Wortfehler (mean) über alle Testsätze für jede Clustervariante in der Anfangsphase (Iteration 36), mittleren Phase (Iteration 50) und Endphase des Trainings (Iteration 56). Die Angaben zur Signifikanz (p) beziehen sich auf die Differenz zwischen den Wortfehlerraten der einzelnen Varianten und der Variante „Cluster 6“. Man erkennt, daß zu Beginn des Trainings größtenteils keine signifikanten Unterschiede (ns) zu der besten Variante „Cluster 6“ bestanden. In der Mittelphase des Training zeichnen sich bereits einige Unterschiede ab und zum Ende des Trainings liegen die Wortfehlerraten aller Varianten signifikant unter den Raten der Variante 6. Außer bei den Varianten 5 und 10 sind die Ergebnisse sogar signifikant auf dem 1% Niveau (**).

7 Ausblick

In der momentanen Situation besteht eines der Probleme bei der akustischen Modellierung von Geräuschen in dem Mangel an Trainingsmaterial. So mußte in dieser Arbeit z.B. auf den Einsatz von Kreuzvalidierungstests verzichtet werden, da ein weiteres Aufteilen und damit Reduzieren der Trainingsdaten nicht vertretbar war. Außerdem mußten viele Geräusche aufgrund zu geringen Vorkommens in Klassen zusammengefaßt werden, was einer differenzierten Modellierung abträglich ist.

Da an vielen Orten inzwischen mit der Sammlung spontan sprachlicher Daten begonnen wurde, wird zukünftig der Mangel an Trainingsmaterial behoben werden können. Die akustische Modellierung von sprachlichen und nichtsprachlichen Geräuschen sollte dann noch auf weitere interessante Aspekte ausgedehnt werden. So konnte beispielsweise festgestellt werden, daß die Länge der einzelnen Geräusche stark variiert. Das Geräusch `+CLICK+` war vielfach sehr kurz, `+UH+` ist ein Beispiel für eine eher langgezogenen Artikulation. Bei den zusammengefaßten Klassen zeigte sich auch innerhalb der Klasse eine starke Variation der Verteilung über die Phonemlänge. Interessant wäre daher zu überprüfen, inwieweit die Längenmodellierung, wie sie für die normalen Phonemmodelle gewählt wurde, auch für die Geräuschmodelle brauchbar sind.
Desweiteren wäre es interessant, die Sprachabhängigkeit von Geräuschen zu untersuchen. Bei Ähnlichkeiten bzw. Gemeinsamkeiten von Geräuschen könnte man eventuell sprachübergreifende Geräuschbibliotheken aufbauen.

Die hier behandelten nichtsprachlichen und sprachlichen Geräusche berühren nur einen kleinen Ausschnitt aus der Problematik spontan gesprochener Sprache. So wurden, um überhaupt einen lauffähigen Erkenner zu erhalten, alle durch Abbrüche, Neuanfänge oder Wiederholungen verstümmelten Wort- bzw. Satzfragmente auf ein gemeinsames Geräuschmodell abgebildet. Eine adäquate Modellierung solcher Phänomene ist eines der vielen Probleme, die zur Steigerung der Erkennungsleistung bei spontan gesprochener Sprache noch gelöst werden muß.

Literatur

[1] J. Butzberger, H. Murveit, E. Shriberg, P. Price: *Modeling Spontaneous Speech Effects In Large Vocabulary Speech Recognition Applications.* SRI, Speech Research and Technologie Programm.

[2] L. Osterholtz, A. McNair, I. Rogina, H. Saito, T. Sloboda, J. Tebelskis, A. Waibel, and M. Woszczyna: *Testing Generality in JANUS: A Multi-Lingual Speech-to-Speech Translation System.* In: Proceedings of the ICASSP 1992, volume 1, pp 209-212.

[3] L. R. Bahl, F. Jelinek, R. L. Mercer: *A Maximum Likelihood Approach To Continous Speech Recognition.* In: A. Waibel and K.-F. Lee, editors, Readings in Speech Recognition 1990, pp 308-319.

[4] R. Schwartz, S. Austin: *A Comparison of Several Approximate Algorithms For Finding Multiple (N-BEST) Sentence Hypotheses.* In: Proceedings of the ICASSP, 1991, pp 701 ff.

[5] O. Schmidbauer: *An LVQ based Reference Model for Speaker-Independent and -Adaptive Speech Recognition.* Internal technical report, Siemens AG, ZFE IS KOM3, 1991.

[6] J.G. Wilpon, L.R. Rabiner, C.H. Lee, E.R. Goldman: *Automatic Recognition of Vocabulary Word Sets in Unconstrained Speech Using Hidden Markov Models.* Transactions ASSP, 1990.

[7] W. Ward: *Modelling Non-verbal Sounds for Speech Recognition* Proceedings of the DARPA Speech and Natural Language Workshop, 1989, pp. 137,141.

[8] W. Ward: *Understanding Spontaneous Speech: The PHOENIX System.* In: Conference Proceedings ICASSP, 1991, pp. 365-368.

[9] C. T. Hemphill, J. J. Godfrey G. R. Doddington: *The ATIS Spoken Language System Pilot Corpus.* Dokumentationsdatei auf der zum ATIS-Task zugehörigen Daten-CD NIST-Disc CD5–1.1, 1990.

[10] M. Woszczyna, N. Aoki-Waibel, F.D. Buø, N. Coccaro, K. Horiguchi, T. Kemp, A. Lavie, A. McNair, T. Polzin, I. Rogina, C.P. Rose, T. Schultz, B. Suhm, M. Tomita, A. Waibel: *JANUS 93: Towards Spontaneous Speech Translation.* To appear in: Proceedings of the ICASSP 1994.

[11] Kai-Fu Lee: *Context-Dependent Phonetic Hidden Markov Models for Speaker-Independent Continuous Speech.* In: IEEE Transactions on Acoustics, Speech, and Signal Processing (ASSP), April 1990.