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

Identifizierung von Sprachen

Exemplarisch aufgezeigt am Beispiel der Sprachen
Deutsch, Englisch und Spanisch

Diplomarbeit
von

Tanja Schultz
(tanja@ira.uka.de)

Erstgutachter: Prof. Dr. Alex Waibel
Betreuer: Dipl.-Inform. Ivica Rogina

Karlsruhe, den 30. 4. 1995

Zusammenfassung

Die automatische Identifizierung von Sprachen gilt als eines der Schlüsselforschungsgebiete, das die Grundlage für die Entwicklung multilingualer Sprachsysteme schafft. Sprachverarbeitende multilinguale automatische Systeme werden angesichts einer zunehmenden Kooperation vieler Partner über Staatsgrenzen hinweg immer dringlicher.

Das Ziel der vorliegenden Diplomarbeit bestand darin, für das multilinguale Übersetzungssystem JANUS ein Modul zu entwickeln, das Sprachen automatisch identifiziert. Die zu entwickelnde Einheit sollte als „front-end Modul" vor JANUS geschaltet werden können und die möglichen Eingabesprachen Deutsch, Englisch oder Spanisch voneinander unterscheiden. Dazu wurden zunächst aus der Literatur bekannte Ansätze als Referenzpunkte nachgebildet und darauf die Leistung des sprachidentifizierenden JANUS-Moduls analysiert. Bei dieser Analyse stellte sich heraus, daß die sehr gute Identifizierung der Sprachen nicht ausschließlich aus der Unterscheidung sprachlicher Merkmale resultierte, sondern vor allen Dingen aus der Erkennung der Kanaleigenschaften des Datenmaterials. Diese entstanden dadurch, daß die Testdaten an unterschiedlichen Orten aufgenommen worden waren. Zur Eliminierung der Einflüsse wurden Cross-channel Daten gesammelt und darauf kanalunabhängige Experimente durchgeführt. Durch die Eliminierung der Kanaleinflüsse verdoppelte sich die Fehlerrate der Sprachenidentifizierung.

Zur Verbesserung der Leistung wurde der Einsatz höherer Wissensquellen zur Identifizierung von Sprachen untersucht. Deren Einfluß war bisher noch nicht Gegenstand der Forschung gewesen. Mehrere neue Ansätze wurden entwickelt und deren Leistung mit den Referenzsystemen verglichen. Durch die Integration höherer Wissensquellen konnte der Identifizierungsfehler im Vergleich zu bisher bekannten sprachidentifizierenden Systeme um die Hälfte reduziert werden.

Inhaltsverzeichnis

1 Problemstellung **1**

2 Identifizierung von Sprachen (LID) **3**
2.1 Bedeutung und Anwendungen der LID 3
2.2 Wissensquellen zur LID 4
2.3 Formulierung des LID-Problems 7

3 LID mit dem JANUS-System **9**
3.1 Ein Überblick 9
3.2 Der Linguistische Decoder 9
3.3 Suche der besten Hypothese 12
3.3.1 Akustische Modellierung mit HMMs 13
3.3.2 Sprachliche Modellierung 15

4 Stand der Forschung zur LID **17**
4.1 Architekturen von LID-Systemen 18
4.2 Klassifikation aktueller Forschungsansätze 20
4.2.1 Akustisch-phonetische Ansätze 21
4.2.2 Phonologische Ansätze 23
4.2.3 Prosodische Ansätze 26
4.2.4 Lexikalische und grammatikalische Ansätze 27
4.3 Ansätze aus der Sprecheridentifizierung 28
4.4 Leistungvergleich sprachidentifizierender Systeme 29

5 Experimente **31**
5.1 Die multilinguale Datenbasis SST 31
5.2 Der traditionelle Ansatz 33
5.3 Die Berechnung der LID-Leistung 37
5.3.1 Normierung durch Subtraktion des Mittelwertes 37
5.3.2 Normierung durch Kalibrieren 39
5.3.3 Berechnung einer Trenngerade mit Neuronalen Netzen 40
5.4 Zeitabhängige Identifizierungsleistung 42
5.5 Das Problem der Kanalabhängigkeit 45

5.6 Kanalunabhängige Experimente 47
5.7 Integration von höherem Wissen 52
5.8 Nachbehandlung durch das Sprachmodell 55
5.9 Grammatikalisch erzwungene Akustik 57
5.10 Die Unterscheidung der Sprachen Deutsch, Englisch und Spanisch 60

6 Diskussion und Ausblick **62**

7 Literatur **63**

Tabellenverzeichnis

4.1 LID-Veröffentlichungen in den wichtigsten Sprachkonferenzen 18
4.2 Identifizierungsleistung aktueller LID-Systeme 30

5.1 Aktueller Stand der Datenbasis SST 32
5.2 Trainings- und Testdatenmaterial 33
5.3 Phonemerkennungsleistung 35
5.4 Geräuschbereinigte Phonemerkennungsleistung 35
5.5 Erkennerabhängige Mittelwerte der Hypothesenscores 38
5.6 LID-Leistungen bei der Normierung durch Subtraktion des Mittelwertes . . 38
5.7 LID-Leistungen bei der Normierung durch Kalibrieren 39
5.8 LID-Leistungen bei Berechnung der Trenngeraden durch Perzeptron 41
5.9 Vergleich der verschiedenen Normierungsverfahren 42
5.10 Erkennerabhängige mittlere Scoresperframe 47
5.11 Datenmaterial für die kanalunabhängigen Experimente 48
5.12 Kanalabhängige mittlere Scoresperframe 48
5.13 Kanalunabhängige LID-Leistung der Systeme PohneLM und PmitLM 51
5.14 Geräuschbereinigte Worterkennungsleistung 53
5.15 Kanalunabhängige LID-Leistung aller Systeme im Vergleich 53
5.16 Kanalabhängige LID-Leistung aller Systeme im Vergleich 54
5.17 LID-Leistungsverbesserungen durch Integration von höherem Wissen . . . 54
5.18 LID-Leistung durch Nachbearbeitung mit Sprachmodell 56
5.19 Wortabhängige LID-Leistung druch Nachbearbeitung mit Sprachmodell . . 57
5.20 LID-Leistung mit CDzwang 58
5.21 LID-Leistung aller Systeme für Deutsch, Englisch und Spanisch 60

Abbildungsverzeichnis

3.1 Das JANUS-System 10

4.1 Architektur eines sprachidentifizierenden Systems 19
4.2 Parallele Architektur mit globalem Ansatz 22
4.3 Phonemmodellierung statt globaler Modellierung 23
4.4 Parallele Architektur mit phonologischem Ansatz 24

5.1 Länge der Testsätze in Sekunden 34
5.2 LID mit akustisch-phonetischem Ansatz 36
5.3 LID mit phonologischem Ansatz 36
5.4 Normierung durch Subtraktion der Mittelwerte 39
5.5 Normierung durch Kalibrieren 40
5.6 Berechnung der Trenngeraden durch Perzeptron 41
5.7 Zeitabhängige LID-Leistung mit und ohne Phonemgrammatik 43
5.8 LID-Leistung inklusive und exklusive SILENCE auf deutschen Sätzen 44
5.9 LID-Leistung inklusive und exklusive SILENCE auf englischen Sätzen 45
5.10 Erkennungsleistung in den ersten 10-100ms 46
5.11 Kanaleinflüsse auf die deutschen und englischen Testdaten 49
5.12 LID-Leistungen am Aufnahmeort CMU 50
5.13 Kanalunabhängige LID-Leistung mit und ohne Grammatik 51
5.14 Nachbehandlung mit Sprachmodell auf CMU Daten 55
5.15 CDpostLM nach Geräuscheliminierung auf CMU Daten 56
5.16 LID-Leistung mit CDzwang am Aufnahmeort CMU 58
5.17 Leistungsvergleich aller vorgestellten Systeme 59
5.18 Unterscheidung Deutsch, Englisch und Spanisch 61

Kapitel 1

Problemstellung

Die wachsende internationale Verflechtung von Wirtschaft, Politik und Gesellschaft erhöht zunehmend den Bedarf an technischen Einrichtungen, die eine schnelle und problemlose Kommunikation von globalem Ausmaß ermöglichen. Da die Zusammenarbeit von Unternehmen über Staatsgrenzen hinweg die Verständigung zwischen verschiedensprachlichen Partnern erfordert, benötigen wir heutzutage Kommunikationssysteme, die Sprachbarrieren überwinden. Dolmetscher, die mehrere Sprachen beherrschen, sind schwer zu finden und teuer. Aufgrund der zunehmenden Informationsmenge und wachsenden Informationsdichte ist ein globaler Informationsaustausch schnell und effizient nur mit automatischen Kommunikationssystemen zu bewältigen. Um vom Anwender akzeptiert zu werden, müssen solche Systeme unkompliziert und problemlos zu bedienen sein und sollten mindestens genauso schnell und zuverlässig funktionieren wie ein menschlicher Dienstleister. Es bieten sich Systeme an, die Dienstleistungen auf der Basis natürlich gesprochener Sprache liefern. Solche Kommunikationssysteme müssen erstens die Sprache verarbeiten, erkennen und verstehen können, zweitens die Fähigkeit der Multilingualität besitzen. Multilinguale Sprachsysteme sind Systeme, die als Ein- und Ausgabesprache verschiedene Sprachen akzeptieren. In vielen Einsatzsituationen ist die Angabe der verwendeten Sprache durch den Anwender beispielsweise anhand eines Knopfdruckverfahrens aus ergonomischen Gründen nicht erwünscht, oder technisch nicht möglich (z.B. am Telefon). Dann ist es notwendig, daß die Sprache, in der gesprochen wurde, vom System selbst identifiziert, d.h. eindeutig von anderen Sprachen unterschieden werden kann. Dies ist die Aufgabe einer sprachidentifizierenden Einheit. Sie soll anhand der Vorgabe eines möglichst kurzen Sprachabschnittes in Echtzeit mit möglichst hoher Präzision die tatsächlich gesprochene Sprache bestimmen.

Die Identifizierung von Sprachen, d.h. die Fähigkeit, Sprachen voneinander unterscheiden zu können, gilt als eine der Schlüsselforschungsgebiete, die eine Grundlage für die Entwicklung multilingualer Sprachsysteme schaffen. Dieses Gebiet ist recht jung und aus theoretischer Sicht sehr interessant, da noch kein Ansatz gefunden wurde, der das Problem der Sprachenidentifizierung befriedigend löst. Die Identifizierung von Sprachen steht in sehr engem Zusammenhang mit der Spracherkennung und der Identifizierung von Sprechern. So haben sich Verbesserungen und Einsichten in einem Bereich auch stets auf die anderen aus-

gewirkt. Dennoch unterscheidet sich die Identifizierung von Sprachen in zwei wesentlichen Punkten von den oben genannten Bereichen: im Gegensatz zur Spracherkennung ist es für die Identifizierung einer Sprache unwichtig, welche Phonem- oder Wortsequenz erkannt wird, und im Gegensatz zur Sprecheridentifizierung sollte die Identifizierung von Sprachen sowohl vom Sprecher als auch von der Domäne, über die gesprochen wird, unabhängig sein.

Das Ziel der vorliegenden Diplomarbeit besteht darin, ein sprachidentifizierendes Modul für das multilinguale Sprachsystem JANUS zu entwickeln. JANUS ist ein Sprache-zu-Sprache Übersetzungssystem, das spontan gesprochene Äußerungen in den Sprachen Deutsch, Englisch und Spanisch erkennt und diese Eingabesprachen wahlweise in die Ausgabesprachen Japanisch, Deutsch, Englisch und Spanisch übersetzt. Die zu entwickelnde Einheit soll als „front-end Modul“ vor das JANUS System geschaltet werden und identifizieren, welche der möglichen Eingabesprachen Deutsch, Englisch oder Spanisch tatsächlich gesprochen wurde. Die Identifizierung soll in möglichst kurzer Zeit mit möglichst hoher Präzision erfolgen. Da die Forschungsarbeiten am JANUS-System fortgesetzt werden und in naher Zukunft als akzeptierte Eingabesprachen Japanisch und Koreanisch sowie als Ausgabesprache Koreanisch hinzukommen werden, ist es wünschenswert, das sprachidentifizierende Modul derart zu gestalten, daß eine Erweiterung möglich ist.

Im zweiten Kapitel werden zunächst die Bedeutung und die Anwendungsgebiete der automatischen Identifizierung von Sprachen dargestellt und die Wissensquellen, die zur Identifizierung herangezogen werden können, erläutert. Daran schließt sich die mathematische Formulierung des Identifizierungsproblems an. Das dritte Kapitel bietet eine knappe Einführung in das multilinguale Sprache-zu-Sprache Übersetzungssystem JANUS und beschreibt die für den Identifizierungsprozeß wichtigsten Module. Im vierten Kapitel wird ein Überblick über moderne Ansätze zur automatischen Identifizierung von Sprachen gegeben und der aktuelle Forschungsstand skizziert. Kapitel fünf beschreibt die eigenen Entwicklungen und die Resultate der Experimente. Das letzte Kapitel faßt die Ergebnisse der Arbeit zusammen und gibt einen Ausblick auf mögliche weiterführende Arbeiten.

Kapitel 2

Identifizierung von Sprachen (LID)

Die automatische Identifizierung von Sprachen (im folgenden abgekürzt mit LID für **L**anguage **ID**entification) bezeichnet den Vorgang der Identifizierung einer vom unbekannten Sprecher gesprochenen Sprache mit Hilfe des Computers.

Die Identifizierung von Sprachen ist eine der schwierigen Aufgaben im Bereich der Spracherkennung und eines der Schlüsselgebiete für multilinguale Sprachsysteme. Die Schwierigkeiten resultieren zum einen aus der großen Variationsbreite der individuellen Sprechweise (Aussprache, Intonation, Sprechgeschwindigkeit, usw.), zum anderen aus der Veränderlichkeit der Themen und Inhalte des Gesprochenen [29]. Je größer nämlich die Varianz innerhalb der gesprochenen Sprache ist, desto stärker werden die Unterschiede zwischen den verschiedenen Sprachen überlagert.

2.1 Bedeutung und Anwendungen der LID

Heutzutage sind sprecherunabhängige und domänenunabhängige Systeme mit großem Vokabular gefordert. Die Akzeptanz von sprachverarbeitenden Technologien beim Benutzer hängt davon ab, wie gut sich diese in bestehende Informations- bzw. Kommunikationssysteme eingliedern lassen. Um die Arbeit eines Benutzers zu erleichtern, muß das System unkompliziert zu bedienen sein. Außerdem müssen die angebotenen Dienste mindestens genauso schnell erbracht werden, wie dies ein Mensch tun könnte. Insgesamt wird immer offensichtlicher, daß viele Portabilitätsvorhaben mehr von der Spezifikation der Domäne und der Ergonomie des Systems abhängen als von den Leistungen der spracherkennenden und -identifizierenden Komponenten selbst [17].

Die Anwendungsgebiete und Einsatzbereiche sprachverarbeitender Systeme sind vielfältig. Denkbar sind alle Arten von Informationssystemen in öffentlichen Institutionen, beispielsweise Zugauskunftsdienste, Messeinformationsstände oder Flughafenservice. Ebenso wünschenswert sind Kommunikationssysteme, die Dienste wie Übersetzung und Vermittlung anbieten. Eine weitere wichtige Anwendung sind telefonbasierte Dienste wie Not-

rufeinrichtungen, medizinische Hilfestellungen, Rufnummernauskünfte, Reisebuchungs- und Hotelreservierungsdienste. Für alle oben genannten Anwendungen ist die Identifizierung der gesprochenen Sprache eine wesentliche Grundvoraussetzung, wenn die entsprechenden Dienstleistungen auch fremdsprachlichen Benutzern angeboten werden sollen. So sind beispielsweise Hotelreservierungen, Terminvereinbarungen und Reisebuchungen für Ausländer, die nicht die landestypische Sprache beherrschen, sehr mühsam und manchmal sogar unmöglich. Für Telefongesellschaften ist es ohne Kenntnis der gesprochenen Sprache schwierig, Anrufe weiterzuvermitteln. Aus ergonomischen und manchmal auch technischen Gründen sollte die Identifizierung von Sprachen automatisch geschehen. Dies würde es erlauben, dem Kunden Dienstleistungen in seiner Sprache anzubieten, bzw. Gespräche an Dolmetscher automatisch weiterzureichen, die die entsprechenden Sprachen beherrschen.

Eine berühmte Einrichtung ist der *Language Line Interpreter* der amerikanischen Telefongesellschaft AT&T im Zusammenhang mit der Notrufnummer 911. Menschen, die in Not oder Panik geraten sind, sprechen häufig in ihrer Muttersprache, auch wenn sie eine andere Sprache in den Grundzügen beherrschen. LID kann hier Leben retten, wenn die Sprache eines hilfesuchenden Anrufers erkannt wird und dadurch schneller Hilfe geleistet werden kann [20].

Weniger dramatisch, aber ebenso nützlich ist der Einsatz eines LID-Moduls als „frontend“ für multilinguale Übersetzungssysteme wie z.B. JANUS. Ein vorgeschaltetes Modul identifiziert die gesprochene Eingabesprache, bevor das Übersetzungssystem die Eingabesprache in die gewünschte Zielsprache übersetzt. Hilfstechniken wie das vorherige Anwählen der gewünschten Eingabesprache, die eine gewisse Vertrautheit mit dem Computer oder die Fähigkeit des Lesens voraussetzen, können dann entfallen.

2.2 Wissensquellen zur LID

Der Mensch ist mit Abstand das beste derzeit verfügbare sprachidentifizierende System. Innerhalb weniger Sekunden kann er entscheiden, ob ihm die gesprochene Sprache bekannt ist. Falls nicht, kann der Mensch zumindest typische Charakteristika der Sprache angeben. Um Computer mit der Fähigkeit der Sprachenidentifizierung ausrüsten zu können, muß zunächst herausgefunden werden, welche Spracheigenschaften bzw. welche Wissensquellen der Mensch zur Identifizierung heranzieht. Erst dann kann entschieden werden, welche Strategien und Informationen für ein automatisches System zweckmäßig sein könnten. Muthusamy et. al. [22] haben zu diesem Zweck eine Untersuchung an je zwei Personen aus zehn verschiedensprachlichen Ländern gemacht. Sie spielten den Testpersonen Sprachauszüge von 1, 2, 4 und 6 Sekunden Länge vor und befragten sie anschließend, welche der zehn möglichen Sprachen gesprochen wurde. Die richtige Antwort wurde den Testpersonen rückgemeldet, so daß sie daraus während der Versuchsreihe lernen konnten. Die Sprachidentifizierungsleistung lag im Mittel bei 69,4% und rangierte zwischen 39,2% und 100%. Als statistisch signifikant erwies sich die Dauer der vorgelegten Sprachauszüge,

die Vertrautheit mit der Sprache und die Anzahl der Sprachen, die den Testpersonen vorher schon bekannt waren. Im Gegensatz dazu lag das beste Ergebnis (Stand 1993) eines automatischen LID-Systems auf denselben Daten bei 55% mit 10 Sekunden dauernden Sprachauszügen[1]. Nach Abschluß der Untersuchungen wurden die Testpersonen nach ihren Strategien zur Identifizierung der Sprachen befragt. Nach ihren Angaben benutzten sie eine Kombination von „Phonem"- und „Wordspotting" sowie die phonetischen und prosodischen Merkmale einer Sprache.

Das Auffinden und Operationalisieren von menschlichen Strategien ist ein wichtiger Schlüssel zur Lösung der automatischen Identifizierung von Sprachen. Prinzipiell kann man zwischen folgenden Informationsquellen, die zur LID herangezogen werden bzw. werden können, unterscheiden:

- akustisch-phonetische Merkmale der Sprache
- phonologische Merkmale der Sprache
- prosodische Merkmale der Sprache
- Regeln über die Bildung von Worten einer Sprache aus Phonemen
- grammatikalische Struktur der Sprache

1. akustisch-phonetische Merkmale der Sprache

Sprachen unterscheiden sich bezüglich des Inventars und der Auftretenshäufigkeit der sogenannten Phoneme, der kleinsten bedeutungsunterscheidenden sprachlichen Einheiten. Jede Sprache enthält einen festen Satz von Phonemen, aus denen alle Worte des Wortschatzes gebildet werden. Die Anzahl und Charakteristik der Phoneme kann von Sprache zu Sprache stark variieren. So enthält beispielsweise die französische Sprache Nasallaute, die im Hochdeutschen selten vorkommen. Das Deutsche hat den typischen velaren Reibelaut /ch/[2] wie in *ich*. Das Englische enthält das /th/ wie in *the*, was einem deutschen Sprecher bekanntlich viele Probleme bereiten kann.

2. phonologische Merkmale der Sprache

Jede Sprache hat ihre individuellen Regeln, auf welche Weise die phonetischen Einheiten kombiniert und koartikuliert werden. Diese Regeln kommen vor allem in der Art und Weise der Wortbildung zum Ausdruck. Die Hawaiianer sind z.B. bekannt für ihre vokalreiche Sprache und den zwischen den Vokalen liegenden glottalen Stoplaut. Im Tschechischen können Laute wie /r/ und /l/ silbenbildend sein, etwa bei *prst (Finger)*, was im Deutschen

[1] siehe auch Kapitel „Leistungsvergleich von Systemen"

[2] Die Schreibweise /phonem/ symbolisiert keine Phoneme im engeren Sinne der Phonologie, sondern ist eine üblicherweise verwendete Umschrift. Wenn also ein englisches /r/ und ein deutsches /r/ erwähnt werden, dann kann es sich durchaus um verschiedene Phoneme mit unterschiedlicher Artikulation handeln

nicht erlaubt ist. Sprachentypisch sind auch Verschleifungen von Wortübergängen und benachbarten Phonemen. Darüberhinaus existieren dialektabhängige Varianten z.B. *samma* im Bayerischen und *simmer* im Badischen für *sind wir*.

3. prosodische Merkmale der Sprache

Betonung, Intonation, Rhythmus, Tempo und Pausen sind weitere wichtige Charakteristika einer Sprache (und auch des Sprechers). Eine Klangsprachen wie z.B. Vietnamesisch, hat eine ganz andere Intonationscharakteristik als beispeilsweise die deutsche Sprache. Reddy [28] sagte einmal, daß die Prosodie in der geprochenen Sprache vergleichbare Bedeutung habe wie die Zwischenräume in der Schrift. Die Tatsache, daß wir Menschen die Prosodie zur Identifizierung einer Sprache heranziehen [22], unterstreicht deren Bedeutung eindringlich.

4. Regeln über erlaubte Wortbildungen in der Sprache

Jede Sprache verfügt über einen eigenen Wortschatz. Zwar entspringen viele Begriffe in verschiedenen Sprachen denselben Wortstämmen, jedoch unterscheiden sich die Wortmuster von Sprache zu Sprache nicht zuletzt durch den sprachtypischen Satz an Phonemen, aus dem die Worte gebildet werden. Für LID-Systeme bedeutet dies insbesondere, daß beispielsweise ein deutschstämmiger Sprecher möglicherweise ein englisches Wort nach deutschem Klangmuster ausspricht, es aber trotzdem als englisch gesprochen erkannt werden soll. Diese Probleme sind nur zu bewältigen, wenn dem sprachidentifizierenden System Wissen auf der Ebene von erlaubten Worten hinzugefügt wird.

5. grammatikalische Struktur der Sprache

Die grammatikalische Struktur eines Satzes kann zwischen verschiedenen Sprachen stark variieren. So sind z.B. die Wortstellungen innerhalb eines Satzes im Deutschen anders als im Englischen. Bei einem Fragesatz befindet sich die zu einem Fragewort gehörende Präposition im Englischen am Satzende, während sie im Deutschen am Satzanfang steht. Aus diesem Grund könnte ein grammatikalisches Sprachmodell bei der Identifizierung der Sprache möglicherweise zusätzlich Aufschluß geben. Dafür muß allerdings ein Spracherkennungssystem bereits vorhanden sein, um den zu identifizierende Sprachausschnitt erkennen zu können.

Ein sprachidentifizierendes System sollte zur Unterscheidung von Sprachen alle diese Informationsquellen einbeziehen.

2.3 Formulierung des LID-Problems

Im folgenden soll eine mathematische Formulierung für das Problem der Identifizierung von Sprachen gegeben werden. Sei $\mathbf{L} = \{L_1, L_2, ..., L_n\}$ die Menge von n zu unterscheidenden Sprachen. Die Aufgabe eines sprachidentifizierenden Moduls besteht darin, zu einer vorgegebenen Testäußerung diejenige Sprache L_i zu bestimmen, die in der Äußerung gesprochen wurde. Die Testäußerung besteht aus einem Auszug gesprochener Sprache, wobei der Sprecher, der Inhalt, die Domäne und natürlich die Sprache, in der gesprochen wurde, unbekannt sind.

Sei $\mathbf{A} = \{a_1, a_2, ..., a_n\}$ die Sequenz der akustischen Signale der Testäußerung. Dann beschreibt $P(L_i|\mathbf{A})$ die bedingte Wahrscheinlichkeit, daß es sich bei dem gegebenen akustischen Signal $\mathbf{A}$ um eine Äußerung in der Sprache L_i handelt.
Der *Maximum Likelihood*-Lösungsansatz basiert auf dem Prinzip, daß diejenige Sprache $\hat{L}$ als die tatsächlich gesprochene Sprache angenommen wird, die bei gegebener Akustik $\mathbf{A}$ die höchste Wahrscheinlichkeit hat. Es gilt also

$$\hat{L} = \underset{L_i}{\operatorname{argmax}}\, P(L_i|\mathbf{A}) \tag{2.1}$$

Da jede gesprochene Äußerung aus einer Sequenz linguistischer Ereignisse besteht, sollte zur Identifizierung der Sprache linguistische Information mit einbezogen werden. Sei $\mathcal{W}$ die Menge aller möglichen linguistischen Sequenzen, die eine Äußerung repräsentieren können. Als Elemente dieser Sequenzen sind ganze Worte, Phonemkategorien (Vokale, Reibelaute ...), Phoneme, Diphone oder gar Triphone gebräuchlich. Zunächst stellt sich die Frage, wie gut eine Sequenz $W \in \mathcal{W}$ mit dem akustischen Ereignis $\mathbf{A}$ übereinstimmt. Bei einem Maximierungsprozeß sucht man deshalb unter allen möglichen Sequenzen diejenige mit der höchsten Wahrscheinlichkeit heraus, d.h.

$$\hat{L} = \underset{L_i}{\operatorname{argmax}} \sum_{W \in \mathcal{W}} P(L_i, W|\mathbf{A}) \tag{2.2}$$

Gleichung 2.2 kann man umformen zu

$$\hat{L} = \underset{L_i}{\operatorname{argmax}} \sum_{W \in \mathcal{W}} P(L_i|W, \mathbf{A}) \cdot P(W|\mathbf{A}) \tag{2.3}$$

Nimmt man nun an, daß die beste passende Sequenz linguistischer Ereignisse $W_{best} = \operatorname{argmax}_{\mathcal{W}} P(W|A)$ unabhängig von der Sprache gefunden werden kann, so reduziert sich die Gleichung 2.3 zu

$$\hat{L} = \underset{L_i}{\operatorname{argmax}}\, P(L_i, W_{best}|\mathbf{A}) \quad , \tag{2.4}$$

was äquivalent ist zu

$$\hat{L} = \underset{L_i}{\operatorname{argmax}}\, P(\mathbf{A}|L_i, W_{best}) \cdot P(W_{best}|L_i) \cdot P(L_i) \tag{2.5}$$

Diese einzelnen Wahrscheinlichkeiten sind einfacher zu modellieren als der komplexe Ausdruck in Gleichung 2.2.

$P(\mathbf{A}|L_i, W_{best})$ bezeichnet man als akustisches Modell, $P(W_{best}|L_i)$ als das sprachliche Model[3] der Sequenz. $P(L_i)$ ist die a-priori Wahrscheinlichkeit für die Sprache L_i, die entfallen kann, wenn man von einer Gleichverteilung der Sprachen in den Testäußerungen ausgeht.

Die Unterschiede, die in Abhängigkeit der gesprochenen Sprache in den akustischen und sprachlichen Modellen auftreten, werden zur Unterscheidung der Sprachen voneinander genutzt.

[3]Im Englischen existieren die verschiedene Begriffe „speech“ und „language“, die im Deutschen im allgemeinen beide mit „Sprache“ übersetzt werden. Mit „speech“ ist immer der Sprechakt an sich gemeint, „language“ bezeichnet die Sprache, in der gesprochen wird, oder das Konstrukt Sprache mit seinen syntaktischen Eigenheiten. Wenn im folgenden vom „sprachlichen Modell“ gesprochen wird, dann ist damit das „language model“ gemeint.

Kapitel 3

LID mit dem JANUS-System

Aufgabenstellung dieser Diplomarbeit war es, für das Übersetzungssystem JANUS ein Modul zu entwickeln, das die gesprochene Sprache unabhängig vom Sprecher erkennt. Dazu ist es zweckmäßig, dieses sprachidentifizierende Modul an die Gesamtarchitektur anzupassen. Im folgenden soll daher zunächst JANUS vorgestellt werden. Dabei wird besonderes Augenmerk auf die Module gelegt, die für die Nutzung von JANUS als LID-System wichtig sind, nämlich das akustische Modell, das sprachliche Modell (language model) und das Suchmodul.

3.1 Ein Überblick

JANUS ist ein Sprache-zu-Sprache Übersetzungssystem, das im Rahmen der Kooperation Interactive System Laboratories zwischen der Universität Karlsruhe am Institut für Logik, Komplexität und Deduktionssysteme und der Carnegie-Mellon University, Pittsburgh entwickelt wird [37].

JANUS ist modular aufgebaut und besteht im wesentlichen aus sieben Bearbeitungsabschnitten. Die Aufteilung in die einzelnen Module und der Verarbeitungsablauf von JANUS ist in Abbildung 3.1 zu sehen. Da für die weitere Vorgehensweise der Arbeit hauptsächlich die Verarbeitungsschritte 3 und 4 von Interesse sind, werden diese im nächsten Kapitel näher beleuchtet. Weitere Ausführungen über die verwendeten Prinzipien in den Bearbeitungsschritten 1,2,5,6 und 7 würden den Rahmen der vorliegenden Arbeit sprengen, der interessierte Leser wird an dieser Stelle auf die entsprechende Literatur verwiesen [37], [38], [35].

3.2 Der Linguistische Decoder

Das Ziel der Verarbeitungsschritte 3 und 4 ist es, aus dem vorverarbeiteten Sprachsignal den Satz zu erkennen, der tatsächlich gesagt wurde. Für kontinuierliche bzw. spontan gesprochene Sprache und besonders für Systeme mit großem Vokabular ist dieses Vorhaben

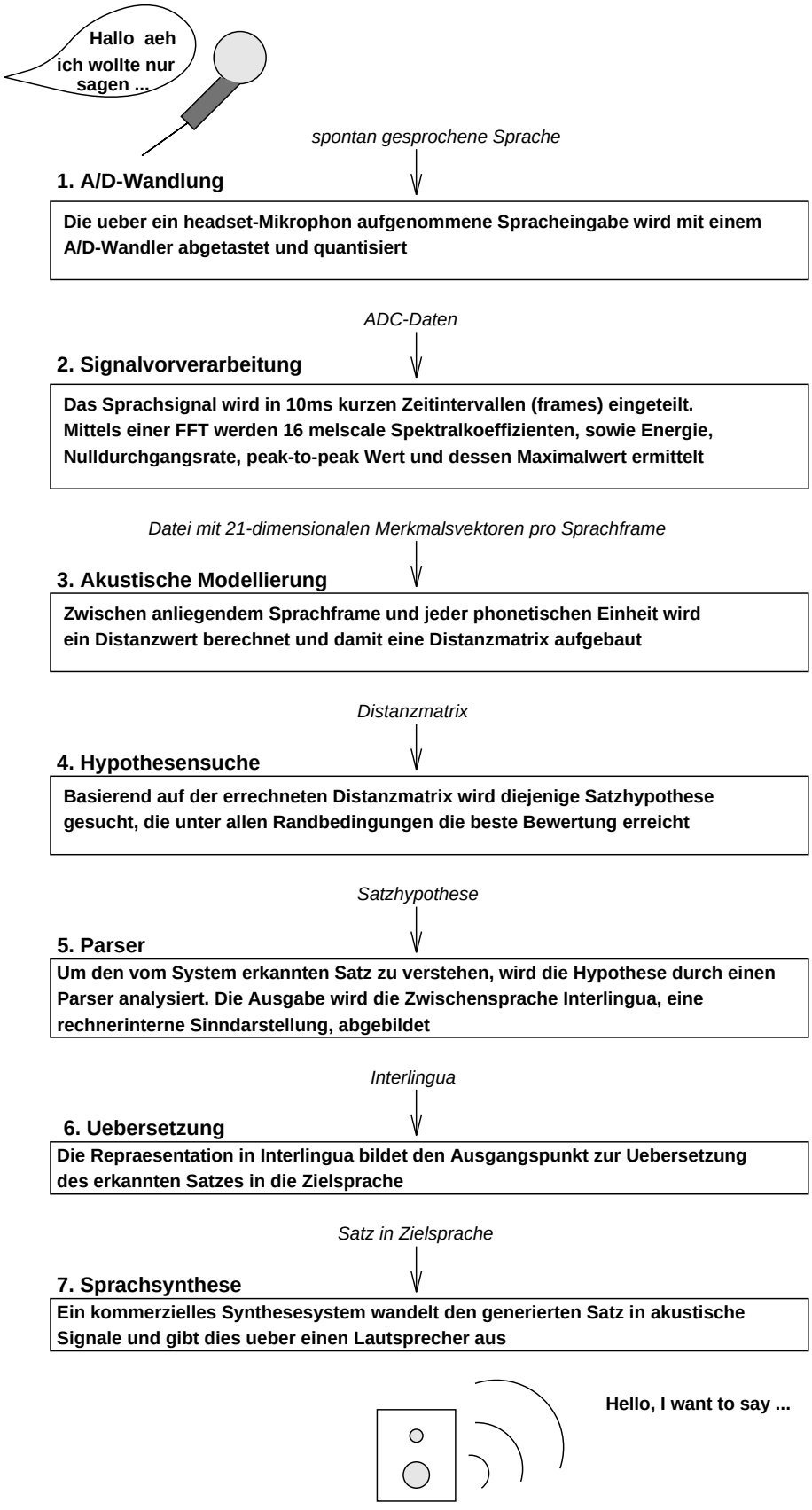

Abbildung 3.1: Verarbeitungsablauf im JANUS-System

in naher Zukunft noch nicht fehlerfrei und zuverlässig möglich. Man wendet daher fehlerkorrigierende Techniken und statistische Methoden an. Dies hat jedoch zur Folge, daß eine große Anzahl von Satzhypothesen entsteht, aus der die geeignete *beste Hypothese* ausgewählt werden muß. Zur Auswahl der besten Hypothese nimmt man Randbedingungen höherer Ordnung zu Hilfe wie sie z.B. durch Syntax, Semantik und Pragmatik gegeben sind [25]. Eine Satzhypothese besteht aus einer Sequenz von Worten. Diese ist genau dann die *beste Hypothese*, wenn sie für das vorliegende akustische Signal am wahrscheinlichsten ist. Die mathematische Formulierung des Problems der Spracherkennung ist dann die folgende [5]:

Sei $\mathbf{A}$ die aus dem akustischen Sprachsignal abgeleitete Beobachtungssequenz, auf deren Basis der Spracherkenner seine Entscheidung darüber fällt, welche Worte gesprochen wurden.

Sei $\mathbf{W}$ eine Wortsequenz bestehend aus n Worten, die aus einem bekannten Vokabular $\mathcal{V}$ stammen.

$$\mathbf{W} = w_1, w_2, \ldots, w_n \quad w_i \in \mathcal{V} \tag{3.1}$$

Der Erkenner entscheidet sich bei gegebener Beobachtungssequenz $\mathbf{A}$ für diejenige Wortsequenz W_{best}, die unter allen Wortsequenzen die wahrscheinlichste ist:

$$W_{best} = \underset{\mathbf{W}}{\operatorname{argmax}}\, P(\mathbf{W}|\mathbf{A}), \tag{3.2}$$

Dabei wird stillschweigend vorausgesetzt, daß alle Worte einer Sequenz inhaltlich die gleiche Wichtigkeit für den Hörer haben, daß somit alle Fehlererkennungen innerhalb der Wortsequenz gleichermaßen gewichtet sind[1].

Mit der Bayes'schen Entscheidungsregel kann man die rechte Seite der Gleichung 3.2 umformen zu

$$P(\mathbf{W}|\mathbf{A}) = \frac{P(\mathbf{W}) \cdot P(\mathbf{A}|\mathbf{W})}{P(\mathbf{A})}, \tag{3.3}$$

wobei $P(\mathbf{A})$ die mittlere Wahrscheinlichkeit ist, daß $\mathbf{A}$ beobachtet wird. Da $\mathbf{A}$ die zugrundeliegende Beobachtung ist, auf der die Wortketten gebildet werden, spielt sie beim Vergleich verschiedener Wortketten untereinander keine Rolle. Damit ergibt sich, daß die beste Hypothese diejenige Wortsequenz W_{best} ist, die das Produkt $P(\mathbf{W}) \cdot P(\mathbf{A}|\mathbf{W})$ maximiert, d.h.

$$W_{best} = \underset{\mathbf{W}}{\operatorname{argmax}}\; P(\mathbf{W}) \cdot P(\mathbf{A}|\mathbf{W}) \tag{3.4}$$

Dieser Ausdruck enthält die Komponenten, die für die Entwicklung eines spracherkennenden Systems auf der Basis stochastischer Modellierung wichtig sind.

$P(\mathbf{W}) = P(w_1 \ldots w_n)$ ist die a-priori Wahrscheinlichkeit, mit der die Wortsequenz $\mathbf{W} = w_1, w_2, \ldots, w_n$ ausgesprochen wird. Das Auffinden der *besten Hypothese* setzt also

[1] Diese Annahme ist sicherlich nicht richtig, denn innerhalb einer Nachricht können bestimmte Worte wichtiger sein als andere (Aufschrei *Feuer* in einem belebten Theater) [14], wollte man aber bei der Modellierung diesen Gesichtspunkt beachten, bräuchte man ein Maß für die Bedeutung eines Wortes im Kontext, eine Entscheidungsinstanz, die den Worten ihre Bedeutung zuweist und ein entsprechend gewichtetes Strafmaß für die Fehlererkennung unterschiedlich bedeutsamer Worte.

voraus, daß man diese a-priori Wahrscheinlichkeit $P(\mathbf{W})$ bestimmen kann. Das ist das schon erwähnte *language model*.

Der Term $P(\mathbf{A}|\mathbf{W})$ beschreibt die *akustische Modellierung*. Er gibt die bedingte Wahrscheinlichkeit dafür an, daß die akustische Sequenz $\mathbf{A}$ beobachtet wird, vorausgesetzt der Sprecher hat die Wortsequenz $\mathbf{W}$ ausgesprochen.

Man erkennt sofort die enge Verwandtschaft zu Gleichung 2.5. Die Lösung des Spracherkennungsproblems ist somit eng verknüpft mit der Lösung des Identifizierungsproblems. Bei der Identifizierung von Sprachen kommt es allerdings nicht auf die erkannte Wortsequenz an, sondern nur auf die Wahrscheinlichkeit, mit der diese Sequenz ermittelt wurde. Im folgenden soll nun erläutert werden, wie die Suche nach der besten Hypothese realisiert wird. Dazu wird eine knappe Einführung in die akustische und sprachliche Modellierung des JANUS-System gegeben.

3.3 Suche der besten Hypothese

Um aus der riesigen Anzahl möglicher Wortsequenzen diejenige auszuwählen, die das Produkt $P(\mathbf{W}) \cdot P(\mathbf{A}|\mathbf{W})$ maximiert, benötigt man eine effiziente Suchstrategie. Dabei soll die eingegebene Sprache optimal auf eine mögliche Sequenz von Worten abgebildet werden.

Für Systeme mit großem Wortschatz wie JANUS werden in der Spracherkennung üblicherweise keine Worte als Erkennungseinheit verwendet, sondern kleinere Untereinheiten wie Silben, Morpheme oder Phoneme. Im JANUS-System werden als Untereinheiten Phoneme benutzt. Ein zu erkennendes Wort wird als Konkatenation dieser Phoneme modelliert. Die Konkatenationsregeln, d.h. die Regeln für die phonetische Umschrift jedes bekannten Wortes werden in einem Wörterbuch festgehalten. Die Umschreibung der Worte geschieht anhand eines festgelegten Phonemsatzes. Die Wortsequenzen bestehen somit selbst wiederum aus einer Sequenz von Phonemen. Zur Suche der besten Wortsequenz benötigt man somit eine akustische Modellierung dieser Phoneme. Man unterscheidet zwei mathematische Grundmethoden zur Modellbeschreibung. Die einen basieren auf dem Prinzip des Mustervergleichs (Pattern Matching oder Template Matching). Diese Methode ist jedoch gerade für kontinuierliche bzw. spontan gesprochene Sprache auf großem Wortschatz nicht geeignet. Der zweite mathematische Typus sind die stochastischen Modelle. Diese basieren auf der Annahme, daß sich der Sprechvorgang mit Hilfe von Wahrscheinlichkeitsprozessen beschreiben läßt. Ein solches stochastisches Modell, das mittlerweile in den meisten erfolgreichen Spracherkennern verwendet wird, ist das *Hidden Markov Modell (HMM)*.

Es gibt zahlreiche und darunter einige sehr gute Abhandlungen über HMMs. Ein sehr empfehlenswertes Tutorial ist das von Rabiner [26]. Eine gute Einführung in semikontinuierliche HMMs befindet sich in [12]. In [31] ist eine sehr gute Darstellung der HMMs

in deutscher Sprache. Dieser kurze Exkurs ist nicht als Tutorial über HMMs gedacht, sondern soll die wichtigsten Begriffe eingeführen, die für das Verständnis der in der LID verwendeten Ansätze benötigt werden.

3.3.1 Akustische Modellierung mit HMMs

Einem HMM liegen zwei verschiedene stochastische Prozesse zugrunde, womit sich neben dem akustischen Ereignis selbst auch dessen zeitliche Verzerrung und Variabilität beschreiben lassen. Die Vorzüge der Modellierung mit HMMs liegen darüber hinaus in sehr speicher- und recheneffizienten Algorithmen. Die Kombination dieser beiden Tatsachen haben zum großen Erfolg des HMM basierten Ansatzes in der Spracherkennung geführt.

Ein Markov Modell besteht aus einer Menge von Zuständen, die durch Zustandsübergänge miteinander verbunden sind. Zwei stochastische Prozesse liegen dem HMM zugrunde. Ein Prozeß, der die Wahrscheinlichkeit beschreibt, ein Symbol aus einem festen Alphabet in einem gegebenen Zustand zu emittieren. Dieser Prozeß ist sichtbar, die ihn beschreibende Wahrscheinlichkeiten bezeichnet man als Emissionswahrscheinlichkeiten. Der zweite Prozeß beschreibt sozusagen die internen Abläufe des HMM innerhalb der Zustandskette. Dieser Prozeß ist nach außen nicht sichtbar, d.h. versteckt, daher der Name *hidden* Markov Modell. Die ihn beschreibenden Übergangswahrscheinlichkeiten geben an, wie wahrscheinlich der Übergang von einem Zustand des HMM in einen anderen ist.

Ein HMM ist durch 5 Komponenten vollständig definiert $\lambda = (N, V, C, B, \pi)$, wobei

1. N die Anzahl der Zustände eines HMMs angibt
2. V die Menge der beobachtbaren Symbole $V = \{v_1, v_2, \ldots, v_M\}$ bezeichnet, die in einem HMM-Zustand emittierbar sind.
   M ist die Kardinalität der Menge V
3. C die $[0, 1]$ $N \times N$-Matrix der Übergangswahrscheinlichkeiten ist, deren Elemente c_{ij} die Wahrscheinlichkeit beschreiben, daß das System zum Zeitpunkt t im Zustand S_j ist, vorausgesetzt daß es zum Zeitpunkt $t-1$ im Zustand S_i war.

$$c_{ij} = P(q_t = S_j | q_{t-1} = S_i) \quad \text{und} \quad \forall i \sum_{j=1}^{N} c_{ij} = 1 \tag{3.5}$$

   Sind in einem HMM die Übergänge von jedem Zustand in jeden anderen erlaubt, dann spricht man von einem *ergodischen* HMM. In der Spracherkennung werden meist *links-rechts* HMMs verwendet. Bei diesem HMM-Typ handelt es sich um eine Zustandskette. Die einzelnen Zustände dieser Kette dürfen nur von links nach rechts durchlaufen werden, somit gilt $c_{ij} = 0,\ j < i$. Das links-rechts HMM erhält einen definierten Anfangs- und Endzustand. Häufig wird die Reichweite der Sprünge von einem Zustand der Folge in einen anderen auf $\Delta = 2$ oder 3 begrenzt, d.h. $c_{ij} = 0$, für $j > i + \Delta$.

4. B die Menge der Emissionswahrscheinlichkeitsverteilungen angibt, wobei

$$b_j(k) = P(a_t = v_k | q_t = S_j), \tag{3.6}$$

somit gibt $b_j(k)$ die Wahrscheinlichkeit an, zum Zeitpunkt t das beobachtete Symbol v_k auszugeben, vorausgesetzt, daß sich das System zum Zeitpunkt t im Zustand j befindet.

5. π eine diskrete Wahrscheinlichkeitsverteilung ist, die für alle Zustände die Wahrscheinlichkeit beschreibt, zum Zeitpunkt $t = 1$ im Zustand i zu sein, $\pi_i = P(q_1 = S_i)$.

Im JANUS-System wird zur Modellierung eines Phonems ein links-rechts HMM verwendet. In jedem Zustand dieses HMMs wird genau ein Symbol v_i emittiert, das in diesem Fall aus einem akustischen Vektor besteht. Geht man von der durchschnittliche Länge eines Phonems von 60 ms aus und berücksichtigt, daß ein Vektor 10ms Sprache beschreibt, so ergeben sich 6 HMM-Zustände zur Modellierung eines Phonems. Durch „Sprünge“ mit einer Reichweite von $\Delta = 2$ und durch die Möglichkeit in einer Schleife in einem HMM-Zustand zu verweilen, wird dieses 6-Zustand-HMM der Längenvariabilität von Phonemen gerecht.

Dieses akustische Modell in Form der HMM wird nun verwendet, um den Ausdruck $P(\mathbf{A}|\mathbf{W})$ zu berechnen. $\mathbf{W}$ ist wie schon gesagt eine Wortsequenz, die durch Konkatenation von Phonemen entsteht und da die Phoneme durch eine Kette von Zuständen modelliert werden, wird $\mathbf{W}$ quasi selbst durch eine Kette von Markovzuständen modelliert.

Bei einem gegebenen HMM-Modell λ berechnet sich die Wahrscheinlichkeit für die Sequenz beobachtbarer Ereignisse $\mathbf{A}$ zu:

$$P(\mathbf{A}|\lambda) = \sum_S \pi_{s_0} \prod_{t=1}^{T} c_{s_{t-1}s_t} b_{s_t}(\mathbf{A}) \tag{3.7}$$

wobei S eine bestimmte Sequenz von Markovzuständen $S \in (s_0, s_1, \ldots, s_t)$ ist, und über alle möglichen Zustandssequenzen des gegebenen Models aufsummiert wird. Bei dem Ausdruck $\mathbf{A} = a_1, a_2, \cdots, a_T$ handelt es sich beim JANUS-System um die Folge der vorverarbeiteten 21-dimensionalen Vektoren, die die ins System hereinkommende Sprache repräsentieren.

Entsprechend der Berechnungsart für die Emissionswahrscheinlichkeiten $b_{s_t}(\mathbf{A})$, unterscheidet man *diskrete* (DHMM), *semikontinuierliche* (SCHMM für SemiContinuousHMM) und *kontinuierliche* (CDHMM für ContinuousDensityHMM) HMMs.

DHMMs werden verwendet, wenn der Beobachtungsraum V diskret ist oder wenn er in eine endliche Menge M von diskreten Beobachtungen quantisiert werden kann[2]. Für diesen diskreten Fall ergibt sich dann eine Wahrscheinlichkeit $b_{s_t}(O_t) = P(O_t = v_k | q_t = S_j)$. Wenn O_t die diskrete Beobachtung zum Zeitpunkt t ist und q_t der Zustand, in dem das System zur Zeit t verweilt. Für den kontinuierlichen Fall handelt es sich um eine kontinuierliche Wahrscheinlichkeitdichte $b_{s_t}(a_t) = p(a_t = v_k | q_t = S_j)$, wobei $\mathbf{A} = a_1, a_2, \cdots, a_T$ eine Sequenz kontinuierlicher Ereignisse ist.

[2]Ein Verfahren, das dies leistet ist die Vektorquantisierung VQ

3.3.2 Sprachliche Modellierung

Nachdem $P(\mathbf{A}|\mathbf{W})$ ermittelt wurde, bleibt vom Ausdruck in Gleichung 3.4 noch zu klären, wie man $P(\mathbf{W})$ berechnet. Bei dieser Zuweisung von Wahrscheinlichkeiten zu den Wortketten kann man verschiedene syntaktische (und semantische) Modelle anwenden. Wenn man keine explizite sprachliche Modellierung verwendet, wird jeder Wortkette die gleiche Wahrscheinlichkeit zugeordnet. Früher benutzte man häufig endliche Zustandsautomaten. Diese sind aber für Systeme mit großen Vokabularen nicht mehr handhabbar. Die heute allgemein übliche Vorgehensweise ist die stochastische Modellierung.

Mit der Bayes Regel kann man den Ausdruck $P(\mathbf{W})$ zerlegen in:

$$P(\mathbf{W}) = \prod_{i=1}^{n} P(w_i|w_1, \ldots, w_{i-1}) \tag{3.8}$$

$P(w_i|w_1, \ldots, w_{i-1})$ ist die Wahrscheinlichkeit, daß das Wort w_i gesagt wird, unter der Voraussetzung, daß vorher schon die Worte $w_1, \ldots, w_{i-1}$ gesagt worden sind. Die Wahrscheinlichkeit für das Auftreten eines Wortes w_i wird somit anhand der kompletten Vorgeschichte $w_1, \ldots, w_{i-1}$ modelliert. Da die Anzahl der möglichen Wortketten mit Länge l auf einem Vokabular, das die Größe $|\mathcal{V}|$ hat, $|\mathcal{V}|^l$ beträgt und dies schnell astronomisch groß wird, aber gleichzeitig die Menge an Daten zur Berechnung der Wahrscheinlichkeiten begrenzt ist, kann man solche Wahrscheinlichkeiten nicht sinnvoll berechnen. Noch wäre es möglich, alle Wahrscheinlichkeitswerte so zu speichern, daß sie bei Bedarf vom Erkenner in akzeptabler Zeit abgerufen werden könnten. Andererseits trifft es auf die Bildung von Sätzen auch meistens nicht zu, daß das i-te Wort von der kompletten Vorgeschichte $w_1, \ldots w_{i-1}$ abhängt. Um sinnvolle Schätzungen für das i-te Wort zu erhalten, bildet man daher für die Vorgeschichte Äquivalenzklassen. Man erhält dann:

$$P(\mathbf{W}) = \prod_{i=1}^{n} P(w_i|\Phi(w_1, \ldots, w_{i-1})) \tag{3.9}$$

Die Art und Weise, wie man diese Äquivalenzklassen Φ bestimmt, und die Wahrscheinlichkeiten schätzt, ist Gegenstand der sprachlichen Modellierung (*Language Modeling*). Bei dem in dieser Arbeit verwendeten JANUS-System werden recht einfache Äquivalenzklassen benutzt: zwei Vorgeschichten eines Wortes w_i sind äquivalent, wenn sie in demselben Wort w_{i-1} enden.

$$P(\mathbf{W}) = \prod_{i=1}^{n} P(w_i|w_{i-1}) \tag{3.10}$$

$P(w_i|w_{i-1})$ bezeichnet man als *Bigramm*[3], die sprachliche Modellierung mit Bigrammen nennt man Bigramm-Grammatik. Entsprechend spricht man von *Trigramm*-Grammatiken, wenn die Vorgeschichten eines Wortes w_i äquivalent sind, falls sie in denselben zwei letzten Worten w_{i-2}, w_{i-1} enden. Die sprachliche Modellierung mit Trigrammen ist mit dem

[3]die Bezeichnung *bigram* stammt aus dem Englischen und wird im folgenden als eingedeutschtes Wort mit deutscher Schreibweise benutzt

JANUS-System ebenfalls möglich. Nach [5] sind Trigramm-Grammatiken die am häufigsten benutzten N-Gramme zur sprachlichen Modellierung. Sie enthalten nach ihren Untersuchungen alle wesentlichen Informationen. Trigramm-Grammatiken haben aber den Nachteil, daß viele Trigramme im Trainingsmaterial nicht vorkommen. Bigramm-Grammatiken genügen den Bedingungen, einerseits einigermaßen ausreichende Informationen über die Vorgeschichte zu enthalten und andererseits Klassen zu bilden, die häufig genug sind, um noch zuverlässig die Wahrscheinlichkeiten abschätzen zu können. Grundlage zur Schätzung dieser Wahrscheinlichkeiten sind Daten von der Art, wie sie der Erkenner produzieren sollte. Wenn also der Erkenner benutzt wird, um Daten rund um die Terminabsprache zu erkennen, dann sollte die Datenmenge zur Schätzung der Wahrscheinlichkeiten auch aus der Domäne Terminabsprachen stammen. In diesen Daten zählt man nun die Häufigkeiten $\mathbf{C}$ des Auftretens von Wort w_i nach dem Wort w_{i-1}. Die Wahrscheinlichkeit des Bigramms wird dann ermittelt durch:

$$P(w_i|w_{i-1}) = \frac{\mathbf{C}(w_i, w_{i-1})}{\mathbf{C}(w_{i-1})} \tag{3.11}$$

Im Test erscheinen manchmal Bigramme, die in den Trainingsdaten nicht vorkamen, und für die daher $\mathbf{C} = 0$ ist. Um auch für solche Fälle eine Schätzung der Wahrscheinlichkeiten zu bekommen, wird die Grammatik geglättet.
Das sprachliche Model ist ein Ausdruck dafür, wie syntaktisch und semantisch wohlgeformte Sätze gebildet werden. Für den Identifizierungsprozeß wird für jede zu unterscheidende Frage eine Grammatik in der oben beschriebenen Form auf dem spracheigenen Vokabular anhand der Trainingsmenge erzeugt.
Insgesamt wird nach der besten Wortsequenz gesucht, die aus einer Folge von Phonemen und daher wiederum aus einer Kette von Markovzuständen besteht. Man versucht somit eine optimale Zustandssequenz zu der gegebenen gesprochenen Äußerung zu finden. Dieser Vorgang wird im Englischen mit *alignment* bezeichnet. Dieses Alignment wird mit einer Technik des *dynamisches Programmierens* realisiert. Für HMMs verwendet man dazu den sogenannten *Viterbi*-Algorithmus.
Für die Erkennung von Sprache interessiert man sich für die vom Viterbi-Algorithmus mit der höchsten Wahrscheinlichkeit erzeugte Sequenz von Worten. Diese beste Satzypothese wird vom Suchmodul an den Parser ausgeliefert. Für die Identifizierung einer Sprache ist es jedoch unerheblich, wie diese Wortsequenz lautet, wichtig ist allein die Wahrscheinlichkeit, mit der sie bewertet wurde. Berechnet man mit einem eigenständigen System für jede zu erkennende Sprache jeweils die beste Hypothese, dann wird diejenige Sprache, die zu dem System gehört, das die beste Bewertung aller besten Hypothesen ermittelt hat, als die gesprochene Sprache identifiziert.

Kapitel 4

Stand der Forschung zur LID

Forschung im Bereich der LID wird seit etwa 20 Jahren betrieben. Die erste bekannte englischsprachige Studie stammt von Atkinson aus dem Jahre 1968 [4], der Intonation und Dauer zur Sprachidentifizierung nutzte. Leonard und Doddington veröffentlichten 1974 einen Bericht über das erste sprachidentifizierende System. Obwohl die LID schon seit langem Gegenstand der Forschung ist, wurden von Muthusamy, der in seiner Doktorarbeit umfangreiche Nachforschungen anstellte, nicht mehr als 14 englischsprachige Puplikationen im Zeitraum von 1974 bis 1992 gefunden [20]. Die Untersuchungen in diesem Zeitraum brachten für die Forschung im Bereich LID aus zwei Gründen wenig Fortschritt:

1. es fehlten vielfach wichtige Hinweise auf experimentelle Details
2. bis zum Jahre 1992 gab es noch keine gemeinsame allgemein zugängliche multilinguale Datenbasis, auf deren Grundlage man die verschiedenen Ansätze hätte evaluieren können.

Im Jahre 1992 wurde vom Oregon Graduate Institute eine multilinguale Datenbasis mit 10 Sprachen in Telefonqualität (OGI-Datenbasis) entwickelt und gesammelt. Die Daten wurden der Öffentlichkeit zugänglich gemacht und im März 1993 wurde die OGI-Datenbasis vom NIST[1] zur Standarddatenbasis für die Evaluation von LID-Algorithmen erklärt[2]. Dieser Umstand führte zu einer Wiederbelebung der LID-Forschung. Nun war es erstmalig möglich geworden, sprachidentifizierende Systeme an standardisierten multilingualen Daten zu evaluieren. Dies führte in der Folgezeit zu einer wahren Explosion von Veröffentlichungen im Forschungsbereich von LID. Das Thema Identifikation von Sprachen wurde erstmalig gemeinsam mit Sprecheridentifikation und Sprechererkennung in eigenständigen Konferenzsitzung innerhalb der wichtigsten Konferenzen zur Sprachverarbeitung ICASSP[3], Eurospeech und ICSLP[4] behandelt.

[1] Nationale Institute of Standard and Technology

[2] Da JANUS nur mit Deutsch, Englisch und Spanisch trainiert wurde und es sich bei der zu identifizierenden Sprache um spontane Mensch-zu-Mensch Äußerungen handeln sollte, wurde die OGI-Datenbasis für diese Diplomarbeit nicht verwendet

[3] International Conference on Acoustics, Speech and Signal Processing

[4] International Conference on Spoken Language Processing

Jahr	Konferenz	# Artikel	eigene Sitzung
1994	ICASSP	7	ja
1994	ICSLP	7	ja
1993	ICASSP	1	ja
1993	Eurospeech	6	ja
1992	ICASSP	0	nein
1992	ICSLP	2	nein
1991	ICASSP	2	nein

Tabelle 4.1: LID-Veröffentlichungen in den wichtigsten Sprachkonferenzen

Insgesamt ist die LID als Forschungsgegenstand noch recht jung und dadurch auch aus theoretischer Sicht sehr interessant. Denn es ist noch nicht erwiesen, welcher Ansatz die besten LID-Leistungen hervorbringt. Da die Anzahl der in den wichtigsten Konferenzen zur Sprachverarbeitung publizierten englischsprachigen Artikeln seit dem Boom 1992 überschaubar ist, soll im folgenden versucht werden, ein möglichst vollständiges Bild der bisherigen Ansätze aufzuzeigen.

4.1 Architekturen von LID-Systemen

Grundsätzlich gibt es zwei verschiedene Architekturen für ein LID-System. In der ersten wird für jede zu unterscheidende Referenzsprache ein eigenständiges Modell trainiert. Bei der Identifizierung laufen alle eigenständigen Modelle parallel und erzeugen beim Dekodieren der unbekannten Testäußerung einen Wahrscheinlichkeitswert oder eine akkumulierte Distanz (je nach Methode). Diejenige Sprache, deren Modell die beste Bewertung für die Testäußerung ermittelt hat, wird als die gesprochene Sprache identifiziert. Diese Architektur wird von einem Großteil der Forscher zugrunde gelegt [17], [10], [39]. Sie soll nachfolgend als *parallele Architektur* bezeichnet werden.
Bei der zweiten Architektur wird ein einziges Modell für alle Referenzsprachen gemeinsam trainiert. In diesem Modell sind also alle sprachspezifischen Eigenheiten integriert. Beim Dekodieren der Testäußerung konkurrieren diese Modelle miteinander. Diese Architektur wird im folgenden *integrale Architektur* genannt. Vertreter dieser Architekturform sind [7], [21].

Nachteil der parallelen Architektur ist, daß mit Ansteigen der Zahl der Sprachen zwar die Gesamtdauer des Identifizierungsprozesses unverändert bleibt, aber der insgesamt zu leistende Rechen- und Speicherbedarf, anwächst. Die Redundanz der geleisteten Rechenarbeit steigt. Nachteil der integralen Architektur ist, daß mit zunehmender Zahl zu identifizierender Sprachen die Anzahl der zu integrierenden sprachspezifischen Einheiten steigt.

Damit wird einerseits das Klassifizierungsproblem wegen wachsender Ambiguitäten schwieriger, andererseits steigt die Berechnungsdauer der Algorithmen an.

Die Grafik 4.1 zeigt die Architektur eines sprachidentifizierenden Systems. Beim parallelen Ansatz verfügt jede Sprache über einen eigenständigen Erkenner. Beim Identifizierungsprozeß laufen diese alle zueinander parallel. Bei der integralen Architektur verschmelzen die Systemteile innerhalb des gestrichelt gezeichneten Rahmens zu einem einzigen gemeinsamen Erkenner. Die nachgeordnete Entscheidung über das beste Modell entfällt dann, da diese Entscheidung ein integrierter Bestandteil des gemeinsamen Erkenners ist.

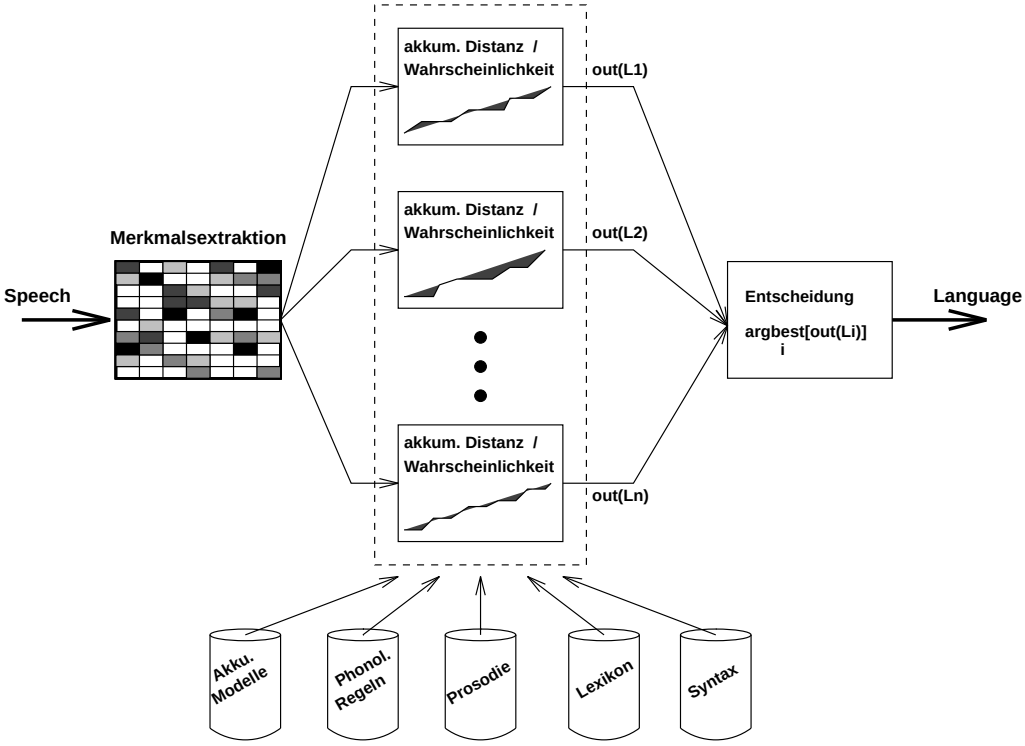

Abbildung 4.1: Architektur eines sprachidentifizierenden Systems

Um die Zahl der sprachspezifischen Phonemmodelle zu verringern, schlugen Dalsgaard, Andersen und Barry [3],[2],[6] ein neuartiges Verfahren vor. Dabei gehen sie von der Annahme aus, daß es in verschiedenen Sprachen Phoneme gibt, die sich akustisch so sehr ähneln, daß man sie gemeinsam modellieren kann[5]. Diese sprachenunabhängigen Phoneme bezeichnen sie als Polyphoneme. Die restlichen Phoneme, die typisch sind für eine einzige, aber nicht für mehrere Sprachen, beispielsweise das bereits erwähnte /ch/ im Deutschen oder das /th/ im Englischen, werden als Monophoneme modelliert. Der Phonemsatz einer Sprache läßt sich dann durch die Vereinigung der sprachentypischen Monophoneme mit der Menge der Polyphoneme bilden. Mit steigender Zahl der zu identifizierenden

[5]Gegenstand ihrer Untersuchungen waren zunächst nur die europäischen Sprachen Englisch, Deutsch, Dänisch und Italienisch

Sprachen reduziert sich durch die gemeinsame Modellierung der Polyphoneme die Anzahl von Modellen und erhöht sich durch die gemeinsame Nutzung das Trainingsmaterial für Polyphoneme. Dalsgaard und Andersen [9] nutzten diese Idee für einen zweistufigen Ansatz mit paralleler Architektur. Sie nahmen an, daß Monophoneme gute Indikatoren für eine spezifische Sprache darstellen. Pro Sprache trainierten sie einen Phonemerkenner, bei dem Mono- und Polyphoneme verwendet werden. Auf der ersten Stufe wird mit Viterbisuche die beste Phonemsequenz generiert. Auf der zweiten Stufe wird auf dieser besten Hypothese ein Nachbearbeitungsschritt durchgeführt. Dabei werden die gegebenen Viterbi-Wahrscheinlichkeiten der Übergänge von einem Phonem in ein anderes mit einem Gewichtungsfaktor multipliziert. Dieser Gewichtungsfaktor entspricht der Verwechslungswahrscheinlichkeit, die jeweils für die Mono- und Polyphoneme auf den Trainingsdaten berechnet wurden. Diejenige Sprache, die mit dem Modell korrespondiert, das in dem beschriebenen Nachmultiplikationsschritt die höchste modifizierte Wahrscheinlichkeit für die Hypothese erzeugt, wird als die gesprochene Sprache identifiziert. Dalsgaard und Andersen legten bisher nur Ergebnisse über sehr wenige Testdaten in 4 Sprachen vor, die noch keine generelle Aussage über den Erfolg des Ansatzes zulassen.

Berkling, Barnard und Arai [7],[8] setzten die Idee der Polyphoneme in der integralen Architektur von Muthusamy et al. [21][6] ein und erreichten damit eine Reduktion der vom neuronalen Netze benötigten Phonembeschreibungsmerkmale um 90% bei nur geringen Leistungseinbußen. Sie stellten aber bei ihren Untersuchungen an sechs Sprachen (Englisch, Deutsch, Spanisch, Japanisch, Hindi und Mandarin Chinesisch) fest, daß es nur sehr wenige Polyphoneme gibt und daher die Einsparungen nicht so hoch waren wie erhofft. Dabei muß man aber berücksichtigen, daß sich die Auswahl ihrer Sprachen nicht nur auf den europäischen Rahmen beschränkte wie bei Dalsgaard, Andersen und Barry.

Leider wurden bei beiden Ansätzen keine Experimente auf dem OGI-Datenbasis veröffentlicht, so daß keine generelle Aussagen über den Erfolg der Ansätze gemacht werden kann. Insgesamt ist noch kein abschließendes Urteil darüber möglich, ob die integrale oder die parallele Architektur für ein sprachidentifizierenden System vorteilhafter ist.

4.2 Klassifikation aktueller Forschungsansätze

Es gibt zahlreiche Möglichkeiten, die bisherigen Ansätze zu unterscheiden. Beispielsweise nach dem verwendeten Datenmaterial (isolierte Worte, gelesene Sprache, spontane Sprache), nach den angewendeten Klassifikationsmethoden (neuronale Netze, HMMs, Vektorquantisierung, Clusteralgorithmen), nach der Anzahl der zu identifizierenden Sprachen (zwei bis elf) oder nach den zur Identifizierung verwendeten Merkmalen der Sprache (akustische Merkmale, prosodische Merkmale, Modellierung differenzierter Phoneme, Modellierung weitgefächerter Phonemklassen, segment- oder silbenbasierte Informationen).

[6] vgl. Abschnitt „Phonologische Ansätze“

Die im folgenden vorgestellten Ansätze werden in dieser Arbeit anhand der Informationsquellen, die sie zur LID heranziehen[7], unterschieden.

Ansätze, die zur Identifizierung einer Sprache nur das akustische Modell heranziehen, sollen nachfolgend als *akustisch-phonetische Ansätze* bezeichnet werden. Ansätze, die bei der Extraktion akustischer Merkmale die Prosodie miteinbeziehen, werden *prosodische Ansätze* genannt. Ansätze, die über die Akustik hinaus das linguistische Modell als Informationsquelle verwenden, werden als *phonologische Ansätze* bezeichnet. Ansätze, die sprachliche Modelle höherer Ordnung wie Lexikon und Grammatik in das linguistische Modell integrieren, werden mit dem Terminus *lexikalische und grammatikalische Ansätze* belegt.

Nachdem die einzelnen Ansätze vorgestellt wurden, sollen im daran anschließenden Kapitel die Ergebnisse der einzelnen Ansätze miteinander verglichen werden, um so den derzeitigen Stand der Forschung zu skizzieren.

4.2.1 Akustisch-phonetische Ansätze

Beim akustisch-phonetischen Ansatz macht man sich die Tatsache zunutze, daß sich Sprachen bezüglich ihrer Kurzzeitakustik unterscheiden. Dies ist nicht nur bedingt durch den Gebrauch unterschiedlicher Phonemsätze, sondern auch durch die unterschiedliche Realisierung gleicher bzw. ähnlicher Phoneme. Als Beispiel für den ersten Fall sei das /ch/ aus dem Deutschen für *ich* genannt, welches im Englischen keine Entsprechung hat. Ein Beispiel für den zweiten Fall ist das /r/, welches sowohl im Amerikanischen Englisch als auch im Deutschen vorkommt. Im Englischen wird das /r/ am Gaumen, unmittelbar hinter den Zähnen gebildet und kontinuierlich ohne Reibung artikuliert. Im Deutschen dagegen wird es je nach Dialekt rollend, schlagend oder reibend ausgesprochen.

Zur Modellierung solcher kurzzeitiger akustischer Effekte hat sich in Spracherkennern die schon beschriebene Methode der Hidden Markov Modelle oder der Neuronalen Netze (NN) durchgesetzt. Die erfolgreichsten Spracherkenner benutzen HMMs zur Modellierung kontextabhängiger Phoneme, die sich als genügend robust gegenüber Sprechervariation und Kontext erwiesen haben. Das Konzept der HMM's hat natürlich auch das Gebiet der LID entscheidend beinflußt.

Seino und Nakagawa [33] zeigten in ihrer Untersuchung, daß eine parallele Architektur, bei der jede Sprache durch ein einziges ergodisches HMM modelliert wird, einem Ansatz basierend auf dem Prinzip der Vektorquantisierung in der Identifikationsleistung überlegen ist. Bei dem ergodischen HMM handelt es sich um ein CDHMM mit 3 Zuständen und einer Gaußverteilung je Zustand. Bei dem Vektorquantisierungsansatz wurde pro Sprache ein Codebook generiert. Die Testäußerung wurde parallel für jede Referenzsprache vektorquantisiert und die Distanzen zwischen Prototypen und anliegendem Sprachframe über die gesamte Äußerung akkumuliert. Das Untersuchungsergebnis veranlaßte Seino und Naka-

[7] wie in Abschnitt „Wissensquellen“ beschrieben handelt es sich um die 5 mögliche Quellen akustisch-phonetische, phonologische und prosodische Merkmale, sowie Lexikon und Grammatik

gawa zu der Vermutung, daß es auch für die Identifizierung von Sprachen eine Rolle spielt, ob die Sprache durch einen gedächtnislosen Prozeß wie die Vektorquantisierung oder durch einen gedächtnisbehafteten Prozeß wie ein HMM modelliert wird.

Bei sprachidentifizierenden Systemen kann man zwischen akustisch-phonetischen Ansätze unterscheiden, die zur Modellierung der Referenzsprachen ein globales HMM oder NN verwenden, und solchen, die von einer differenzierteren Modellierung in Form von Phonemkategorien (broad phonemes) oder Phonemen (fine phonemes) Gebrauch machen.

Zissmann [39] trainierte in einer parallelen Architektur für jede Sprache ein Ein-Zustand-HMM mit 40 gebundenen Gauß'schen Wahrscheinlichkeitsdichten und verglich diese Vorgehensweise mit Mehr-Zustand-HMMs. Die besseren Ergebnisse erzielte das Ein-Zustand-HMM. Zissmann selbst räumte aber ein, daß kein ausreichendes Datenmaterial vorhanden war, um die Parameter des Mehr-Zustand-HMM's zu trainieren. Muthusamy et al. [21] trainierten in einer integralen Architektur ein einziges vollständig verbundenes neuronales „feed-forward" Netz mit einem Backpropagation-Algorithmus auf der Basis roher akustischer Merkmale (PLP-Koeffizienten 7. Ordnung). Das NN klassifizierte jeden Sprachframe als zu einer von 2 Sprachen gehörend (Englisch oder Japanisch). Die Ausgangsaktivierungen für jede Sprache über die Sprachframes wurden über die gesamte Äußerung akkumuliert. Die Sprache mit dem höchsten akkumulierten Wert wurde als die tatsächlich gesprochene Sprache angenommen. Die Grafik 4.2 zeigt den beschriebenen globalen Ansatz als parallele Architektur, bei dem für jede zu erkennende Sprache L_i ein ergodisches HMM eingesetzt wird wie bei Zissmann [39] beschrieben.

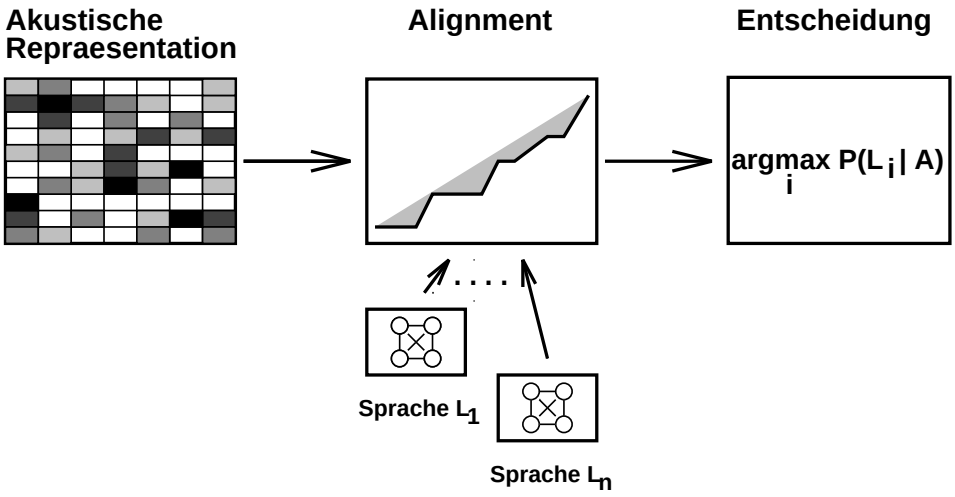

Abbildung 4.2: Parallele Architektur mit globalem Ansatz

Muthusamy et al. verglichen diesen Ansatz mit feinerer Modellierung akustischer Merkmale in Form von Phonemen und Phonemkategorien und stellten fest, daß ein einzelnes NN für die LID keine ausreichende Differenzierung der Sprachen gewährleistet. Zwar hat

der Ansatz eines globalen HMM oder NN pro Sprache den Vorteil, daß keine transkribierten Daten benötigt werden, was den Sammelprozeß der Daten beschleunigt und erheblich kostengünstiger gestaltet, aber insgesamt muß der Ansatz aufgrund seiner niedrigen Identifizierungsleistung als nicht erfolgreich eingestuft werden. Globale Netze bzw. Ein-Zustand-HMMs scheinen nicht geeignet, die Komplexität einer Sprache befriedigend widerzuspiegeln.

Die Wissenschaftler gingen daher zur Modellierung einzelner Phoneme bzw. größerer Klassen von Phonemkategorien jeder Sprache durch eigenständige HMM über. Viele der Forscher sprechen sich dabei für die Modellierung von Phonemen als die geeignetste Methode aus [17],[21],[39]. Bei der feineren Modellierung kommen eher links-rechts HMMs zur Anwendung, weil sie die zeitliche Struktur der Sprache besser nachbilden. Die Abbildung 4.3 zeigt, wie für jede Sprache die ergodischen HMMs auf der linken Seite durch einzelne links-rechts HMMs für jedes Phonem ersetzt werden.

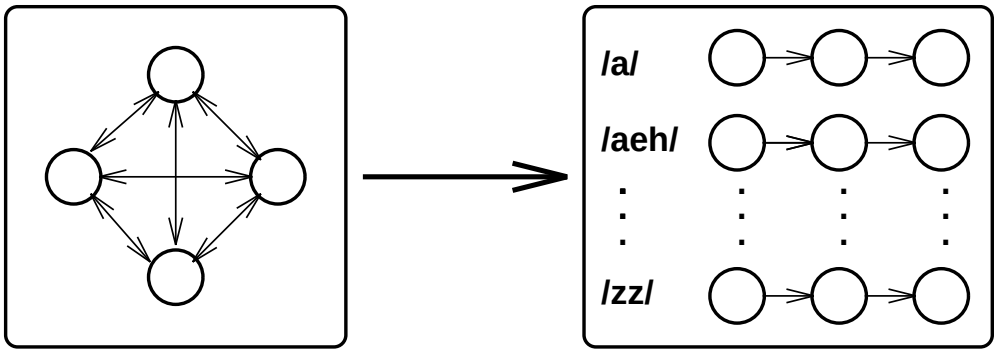

Abbildung 4.3: Phonemmodellierung statt globaler Modellierung

Zissmann und Singer [40] und Muthusamy et al. [21] führten Untersuchungen durch, bei denen jedes Phonem eigenständig modelliert (fine phonemes) wird und verglichen die LID-Leistungen mit Modellierungen von Phonemkategorien wie etwa Vokal, Klanglaut, Reibelauten, Stop und Pause (broad phonemes). Beide Untersuchungen kamen getrennt zu dem Ergebnis, daß die Leistung der Systeme mit zunehmender Differenziertheit der Modellierung anstieg. Dieser akustisch-phonetische Ansatz mit feiner Phonemmodellierung wird meist in Verbindung mit phonologischen Ansätzen verwendet.

4.2.2 Phonologische Ansätze

Phonologische Ansätze sind Verfahren, bei denen zur Identifizierung der Sprache linguistisches Wissen in den Dekodierungsprozeß eingebunden wird. Der Einsatz linguistischen Wissens beschränkt sich aber auf Einheiten wie Phoneme oder Kategorien derselben. Manche System nutzen lediglich die Kenntnis über die Häufigkeit gesehener Phoneme, meist kommen jedoch stochastische Modelle zum Einsatz. Dabei handelt es sich um Phonemgrammatiken, mit Uni- und Bigrammen, seltener auch Trigrammen. Auf diese Weise modelliert man sprachentypische phonotaktische Regeln, also Vorschriften, nach denen in einer

Sprache Phoneme aneinandergefügt werden dürfen. So kann beispielsweise im Englischen auf ein /s/ ein /w/ folgen (*sweat, switch*) wogegen diese Kombination im Deutschen nicht vorkommt[8]. Der Phonemübergang von /s/ nach /w/ bekäme daher im Englischen eine hohe Wahrscheinlichkeit zugeordnet, im Deutschen eine sehr niedrige. Die Abbildung 4.4 zeigt den phonologischen Ansatz für die parallele Architektur mit HMMs.

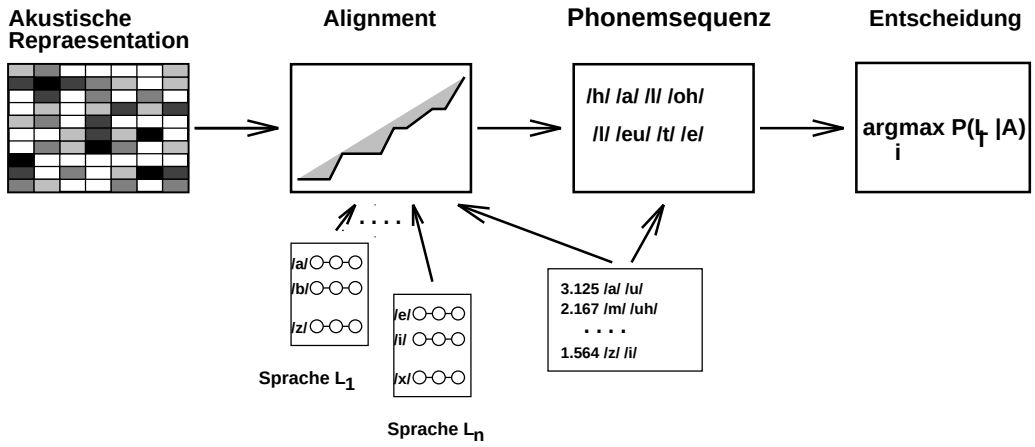

Abbildung 4.4: Parallele Architektur mit phonologischem Ansatz

Die Identifizierung von Sprachen mit einem solchen System sieht folgendermaßen aus: zunächst werden wie beim akustisch-phonetischen Ansatz Phonemmodelle trainiert. Diese werden zur Schätzung der Phonemgrammatik für jede Sprache benutzt. Die unbekannte Testäußerung wird dekodiert und bekommt pro Sprachmodell einen Wahrscheinlichkeitswert zugeordnet.

Der Ansatz von Lamel und Gauvain [17],[18] geht auf die Ideen von House und Neuberg (1977) zurück und wird von ihnen als *phoneme-based acoustic likelihood approach* bezeichnet. Dabei handelt es sich um eine parallele Architektur mit dem Maximum Likelihood Lösungsansatz, bei der für jede zu identifizierende Sprache ein Satz phonembasierter ergodischer HMMs trainiert wird. Bei ihrer Implementierung besteht jedes große ergodische HMM aus kleinen 3-Zustand links-nach-rechts HMMs für die einzelnen Phoneme. Der Wahrscheinlichkeitswert wird mit dem Viterbi-Algorithmus berechnet. Die Phonemgrammatik in Form von Phonembigrammen ist ein integraler Bestandteil des Dekodierungsprozesses. Zissmann [39] unternahm differenzierte Studien darüber, welche Vorgehensweise bei der Einbeziehung einer Phonemgrammatik vorzuziehen sei, die integrale oder die „post-processing“ Variante. Dazu verglich er die folgenden zwei Varianten miteinander: Eine parallele Architektur auf der Basis von Phonemerkennung mit in den Dekodierungsprozeß integrierter Phonemgrammatik (wie bei Lamel und Gauvain). Die Interphonem Übergangswahrscheinlichkeit zwischen zwei Phonemmodellen i und j ergibt sich zu $a_{ij} = s \log P(j|i)$

[8] im Duden befinden sich die drei Einträge *swing, swimming pool* und *sweater*, die alle aus dem Englischen übernommen wurden

wobei s der Grammatik-Skalierungsfaktor und die $P(j|i)$ die interpolierten, vom Training abgeleiteten, Bigramm-Wahrscheinlichkeiten repräsentieren. Den Skalierungsfaktor stellte Zissmann anhand einer Crossvalidierungsmenge von Hand ein. Variante 2 ist mit der ersten bis auf die Tatsache identisch, daß die Phonemgrammatik nicht in den Dekodierungsprozeß integriert ist, sondern als Nachverarbeitungsschritt auf die beste Phonemsequenz angewendet wird. Somit generiert der Phonemerkenner zunächst für eine Sprache die Phonemsequenz der Testäußerung, und ermittelt auf dieser Phonemsequenz in einem nachfolgenden Schritt den Wahrscheinlichkeitswert für jede sprachspezifische Grammatik. Die interpolierte Grammatik λ_L^{BG} der Sprache L produziert somit für die Phonemsequenz W den Wahrscheinlichkeitswert $lP(W|\lambda_L^{BG}) = \sum_{t=1}^{T} \log P(w_t|w_{t-1}, \lambda_L^{BG})$, und es wird diejenige Sprache identifiziert, deren logarithmische Wahrscheinlichkeit maximal ist $\hat{l} = \text{argmax}_L\, lP(W|\lambda_L^{BG})$. Zissmann stellte fest, daß die Leistungen beider Systeme nahezu identisch sind. In seinen Experimenten scheint es also unerheblich, ob die Grammatik in den Dekodierungsprozeß integriert ist oder nicht.

Tucker, Carey und Parris [36] wendeten auf die beste Hypothese, die vom Viterbi-Algorithmus ermittelt wurde ebenfalls eine postprocessing Variante an. Im Nachberechnungsschritt gewichteten sie die aus der Hypothese gesehenen Phonemauftrittshäufigkeiten mit den aus dem Training geschätzten sprachspezifischen Auftrittswahrscheinlichkeiten.

Muthusamy et al. [21] kombinierten die „postprocessing“ Variante mit einem voll verbundenen neuronalen „feed forward“-Netze mit Backpropagation. Dabei verglichen sie ein NN, das im ersten Durchlauf sieben weitgefächerte Phonemkategorien[9] klassifizierte mit einem NN, das feiner differenzierte Phoneme (39 englische und 25 japanische Phoneme) klassifiziert. Im ersten Schritt entscheidet das NN für jeden Sprachframe, um welche Kategorie bzw. Phonem es sich handelt. Im zweiten Schritt wurde darauf aufbauend anhand von Unigramm Eigenschaften und Bigramm-Übergangswahrscheinlichkeiten aus der Phonemsequenz die Sprache bestimmt. Muthusamy et al. stellten fest, daß das NN zwar mit feiner werdender Modellierung der akustischen Ereignisse diese immer schlechter klassifizierte, daß sich aber gleichzeitig die LID-Leistung des Systems verbesserte. Offensichtlich wird die schlechtere Klassifizierungsleistung durch die differenziertere Information, die zum Identifizieren der Sprache zur Verfügung gestellt wird, mehr als kompensiert. Diese Ergebnisse korrespondieren mit denen von Zissmann und Singer [40], die vergleichbare Experimente mit HMMs durchführten.

Meist werden Phonembigramme zur Modellierung der phonologischen Regeln benutzt. In drei Arbeiten von Kadambe/Hieronymus, von Hazen/Zue und von Reyes/Seino/Nakagawa wurden zur grammatikalischen Modellierung der Phoneme Trigramme verwendet. In der Arbeit von Kadambe und Hieronymus [15] brachte die Verwendung von Phonemtrigrammen eine signifikante Verbesserung gegenüber keiner expliziten grammatikalischen Modellierung. Zum Training ihrer Trigramme benutzten sie ein 10-Millionen-Worte-Datenbasis,

[9] Vokal, Reibelaut, Stop, prä-,inter-,postvokaler Klanglaut und Silence

die aus Zeitungen gewonnen wurde. Sie beobachteten, daß die Verbesserungen größer waren bei Sprachen deren akustische Merkmale ähnlich sind (Englisch-Spanisch), als bei unähnlichen Sprachen (Englisch-Mandarin Chinesisch).

Hazen und Zue [10] verglichen in ihrer Untersuchung verschiedene Grammatiken, die sie ausschließlich auf dem OGI-Datensatz erstellten. Für Sequenzen von breiten Phonemkategorien stellten sie fest, daß die Bigramm-Modellierung sowohl der Unigramm-, als auch der Trigramm-Modellierung überlegen war. Dies mag daran liegen, daß Bigramme eine stärkere Bedingung darstellen als Unigramme, andererseits aber für Trigramm-Modellierung meist nicht genügend Datenmaterial zur Verfügung steht, um die Trigramm-Wahrscheinlichkeiten zuverlässig zu schätzen. Das Fehlen von relevantem Trainingsmaterial ist zugleich das Hauptproblem der Modellierung mit Trigrammen.

Reyes, Seino und Nakagawa [29] konnten in ihrer Studie zeigen, daß die Anwendung von Trigrammen erst für Äußerungen ab einer bestimmten Mindestlänge (10 Sekunden) wirklich greift. Je länger die Äußerung, desto größer der Zugewinn durch die Grammatik.

4.2.3 Prosodische Ansätze

Die prosodische Information wie etwa Grundfrequenz, Amplitude und Geschwindigkeit ist ein weiteres wesentliches Charakteristikum von Sprache. Obwohl die Prosodie eine wichtige Informationsquelle für die Spracherkennung ist, haben aktuelle Spracherkennungssysteme noch keinen signifikanten Gewinn aus dem Einsatz prosodischer Information erzielen können. Auch für den Bereich der LID läßt der durchschlagende Erfolg noch auf sich warten. Bei der Identifizierung von Sprachen wurde die Prosodie als Informationsquelle bisher nur in zwei Ansätzen veröffentlicht.

Die Wissenschaftler Itahashi, Zhou und Tanaka [13] gingen von der Tatsache aus, daß der Mensch in einer geräuschbehafteten Umgebung (z.B. fahrender Zug) häufig noch den Dialekt oder die Intonation identifizieren kann, obwohl er das eigentlich Gesagte nicht mehr erkennt. Dies bekräftigt ihrer Meinung nach die große Bedeutung der Prosodie und anderer segmentaler Information, also Langzeitakustik für die Identifizierung der Sprache. Sie benutzten daher als Basis ihres LID-Systems die Grundfrequenz als ein prosodisches Merkmal und approximierten die Grundfrequenzkontur von Sprache über die gesamte Testäußerung durch eine Menge von Geraden. Die Autoren optimierten die Anzahl der Geraden, indem sie mit einer minimalen Menge möglichst globale Merkmale zu beschreiben versuchten. Aus den errechneten Geraden extrahierten sie 17 Merkmale wie Startfrequenz, Steigung und Dauer jeder Geradenkomponente, sowie Maße wie Mittelwerte und Standardabweichungen, und Energie. Die Anzahl der Merkmale reduzierten sie durch die Anwendung einer Faktorenanalyse. Mit den verbleibenden relevanten Merkmalen führten sie eine Diskriminanzanalyse durch und kamen so zu einer guten Diskrimination von 6 Sprachen (Japanisch, Koreanisch, Chinesisch, Englisch, Französisch und Deutsch). Leider verwendeten Itahashi et al. keine übliche Datenbasis, sondern eine Eigenentwicklung mit sehr wenig Sprachmaterial. Daher ist ein objektiver Vergleich mit anderen Verfahren nicht möglich.

Hazen und Zue [10] statteten ihr sprachidentifizierendes System mit dem Wissen über die Grundfrequenz als zusätzliche Komponente aus, und stellten fest, daß diese Information nach der akustischen Information und der Sprachmodellierung anhand von Phonembigrammen den geringsten Einfluß auf die Leistungsverbesserungen des Systems hatte.

Obwohl die Verbesserungen durch den Einsatz prosodischer Information bisher weit hinter dem zurückblieben, was man sich erhoffte, gibt es viele Gründe anzunehmen, daß Prosodie in zukünftigen Systemen eine wichtige Rolle spielen wird. So zeigte bereits die erwähnte Studie von Muthusamy [22], daß die Prosodie für den Menschen als Quelle der Information für die Identifizierung von Sprachen sehr wichtig ist.

4.2.4 Lexikalische und grammatikalische Ansätze

Unter allen analysierten Forschungsansätzen konnte nur die Arbeit von Ramesh und Roe gefunden werden, die Wissen über ganze Worte in den sprachidentifizierenden Prozeß miteinbeziehen. Muthusamy et al. [22] hatten bereits festgestellt, daß der Mensch bei der Identifizierung einer Sprache Techniken wie „Wordspotting'" heranzieht.

Ramesh und Roe [27] kombinierten einen parallelen CDHMM-Ansatz mit einem „Wordspotting"-Verfahren. Für jede Referenzsprache wählten sie dazu etwa 30-40 Schlüsselwörter, die mit links-rechts HMMs von 5-10 Zuständen als ganze Worte modelliert werden. Die eingegebene Sprache wird anhand der sprachentypischen Modelle inklusive Schlüsselwörter (und Hintergrundmodell) dekodiert. Die *Maximum Likelihood*-Entscheidungsregel bestimmt die gesprochene Sprache. Die Autoren erreichten mit diesem an „Wordspotting" angelehnten Verfahren eine signifikante Verbesserung, allerdings nur bei denjenigen Testäußerungen, die auch eines der abgespeicherten Schlüsselworte enthielten. Ihr Ansatz hängt somit von der Wahl geeigneter Schlüsselworte und damit auch von der Domäne ab. Eine Nebenbedingung der LID, nämlich die Idenfizierung der Sprache ohne Einschränkung der Domäne wird damit verletzt. Ramesh und Roe bezeichnen ihren Ansatz denn auch mit *restricted-domain approach*. Solche Verfahren können aber dennoch in Anwendungen mit beschränktem Wortschatz (z.B. Telefonbanking) erfolgreich eingesetzt werden.

Auffallend ist, daß in keiner der zum Thema LID gefundenen Arbeiten die höher geordneten Wissensquellen *Vokabular* und *syntaktische bzw. grammatikalische Regeln* ausgenutzt werden. Es wird die Meinung vertreten, daß phonembasierte Ansätze effizienter sind als wortbasierte Ansätze [17] bzw. daß für multilinguale Ansätze die Einbeziehung höhergeordneter Wissensquellen zu zeit- bzw. rechenaufwendig sei [11]. Diese Einwände sind sicherlich berechtigt, wenn es sich bei dem anvisierten System um ein ausschließliches sprachidentifizierendes Systems handelt. Wird allerdings ein LID-Modul als „front-end" eines Übersetzungssystems eingesetzt, treffen die Kritikpunkte nur noch bedingt zu. Wissensquellen wie Wörterbuch und syntaktische Regeln (Language Models) müssen zur Erkennung ohnehin erzeugt werden. Es liegen dann auch transkribierte Daten vor, so daß kein zusätzlicher

Zeit- und Kostenaufwand für die Erstellung einer Datenbasis entsteht [10]. Der zusätzliche Rechenaufwand, der beim Dekodierungsprozeß, auf Wortebene entsteht, könnte für die Erkennung ausgenutzt werden. Vor allem aber ist das Wissen über erlaubte Wortbildungen dann wichtig, wenn ein Sprecher nicht in seiner Muttersprache spricht, sondern in einer Fremdsprache, und dennoch überwiegend Phoneme seiner Muttersprache verwendet. Das akustisch-phonetische und phonologische Wissen reicht für die Identifizierung solcher mit ausländischem Akzent geprochener Sprachausschnitte nicht aus.

4.3 Ansätze aus der Sprecheridentifizierung

Wie bereits erwähnt konnte die LID in mehrfacher Hinsicht aus Entwicklungen und Verbesserungen in den verwandten Forschungsgebieten Sprecheridentifizierung und Spracherkennung profitieren. Die Überschneidungen zwischen den Gebieten sind sehr groß und so stellten sich einige Wissenschaftler die Frage, ob LID überhaupt von diesen zwei Gebieten entkoppelbar ist. Abe et al. [1] stellten in einer Untersuchung an bilingualen Sprechern fest, daß auf der akustischen Ebene die Unterschiede zwischen den Sprechern größer sind als zwischen den Sprachen. Bei der Identifizierung von Sprachen hat man mit Unterschieden zu tun, die resultieren können aus:

1. Text bzw. Domäne
2. Sprecher
3. Umwelt und Kanaleigenschaften
4. Sprache

Li [19] ging der Frage nach, wie und ob überhaupt sprachunterscheidende Merkmale auf akustischer Ebene von den restlichen Merkmalen trennbar sind. Die bisher skizzierten Ansätze legten alle die Annahme zugrunde, daß Modelle für Sprachen sprecherunabhängig trainiert werden können. Das muß aber nicht richtig sein.

Li entwickelte ein System, das Sprachen mit Hilfe der Sprecheridentifizierung bestimmt. Er ordnet eine Testäußerung einer Sprache zu, indem er die Ähnlichkeit zwischen dem Sprecher der Äußerung und allen (abgespeicherten) Sprechern der Referenzsprachen bestimmt. Der Autor nutzt folglich die Ähnlichkeit der Sprechercharakteristika als Indikator für die Sprache. Für das Training benutzt er ein neuronales Netz, um Silbenkerne der Trainingsäußerungen zu extrahieren. Anschließend werden Spektralkoeffizienten an verschiedenen Stellen innerhalb des Kernes extrahiert und gespeichert. Während des Indentifizierungsprozesses werden von der Testäußerung in derselben Weise Silbenkerne extrahiert und deren spektrale Koeffizienten mit für jeden Sprecher gespeicherten Koeffizienten verglichen. Die kleinste Distanz zwischen jedem Kern der zu klassifizierenden Testäußerung

[10] Dies hat eine signifikante Verbesserungen der LID-Leistungen zur Folge, siehe Kapitel „Leistungsvergleich“

und allen gespeicherten Silbenkernen für jeden Sprecher wird berechnet. Die Summe der Differenzen bildet die Distanz zwischen Äußerung und Referenzsprecher. Zur Identifizierung des Sprechers wird dann der Sprecher mit der minimalen Distanz bestimmt. Zur Identifizierung der Sprache wird diejenige gewählt, deren Mittelwert der M ähnlichsten Sprecher minimal wird. Diese Vorgehensweise setzt eine möglichst repräsentative Auswahl von Sprechern jeder Sprache voraus[11].

Obwohl viele Ansätze zur LID Ideen und Algorithmen aus der Sprecheridentifizierung entliehen haben, ist der Ansatz von Li der einzige, der Sprecherähnlichkeiten direkt als Indikator zur LID ausnutzt.

4.4 Leistungvergleich sprachidentifizierender Systeme

In vorangehenden Abschnitt wurde versucht, einen Abriß über die aktuellen Forschungsansätze zu geben. Da die Ideen der Ansätze vorgestellt werden sollten, wurde von vielen Details, wie etwa die Art der Vorverarbeitung, abstrahiert. Abschließend sollen nun die vorgestellten Ansätze bezüglich ihrer Leistungsfähigkeit miteinander verglichen werden.

Um einen objektiven Vergleich zu ermöglichen, konnten nur diejenigen Systeme aufgenommen werden, die Ergebnisse auf der Evaluationsdatenbasis OGI veröffentlicht haben. Der übliche OGI-Evaluationstest ist ein 10 Sprachen Test. Hierbei wird dem System eine Testäußerung unbekannter Sprache vorgelegt und das System entscheidet sich für eine von 10 möglichen Sprachen. Die Trefferquote wird prozentual ermittelt. Bei den Testäußerungen handelt es sich wahlweise um 10 Sekunden dauernde Sprachauszüge oder komplette Äußerungen unterschiedlicher Länge.

Trotz fest definierter Testmenge ist vom direkten Vergleich der prozentualen LID-Leistung schwer zu sagen, welcher Ansatz den übrigen vorzuziehen sei, da in den Veröffentlichungen nicht immer alle Details über zusätzliche Maßnahmen wie Kanaladaption oder andere Faktoren, die erheblichen Einfluß auf die Leistungen haben können, erläutert werden.

[11] womit sich die Frage stellt, was repräsentativ bedeutet

Autoren, Quelle	Gesamt	10 Sekunden	Methode
nach 1994 - mit transkribierten Daten			
Zissmann/Singer, [40]	79,2%	63%	HMM pro Phonem+Bigramm
Li, [19]	78%	59%	sprecherbasierte LID
Hazen/Zue, [11]	70,1%	65,4%	Phoneme+Bigramm+Prosodie
vor 1994 - ohne transkribierte Daten			
Muthusamy, [23]	66%	47,7%	Neuronale Netze
Lamel/Gauvain, [17]	-	59,7%	HMM pro Phonem+Bigramm
Hazen/Zue, [10]	57%	47,7%	Phonem+Bigramm+Prosodie
Zissmann, [39]	-	46%	1-Zustand-HMM

Tabelle 4.2: Identifizierungsleistung aktueller LID-Systeme

Die Tabelle 4.2 ist in zwei Abschnitte vor und nach 1994 unterteilt. Vor 1994 standen noch keine transkribierten Daten für die OGI-Datenbasis zur Verfügung, somit ist die Ausgangsbasis der Ansätze vor und nach 1994 verschieden. Aus der Tabelle lassen sich einige Schlüsse für die weitere Forschungsarbeit im Bereich LID ableiten:

1. Das Trainieren auf transkribierten Daten führt zu signifikanten Verbesserungen bei der Identifizierung von Sprachen.

2. Die Identifizierungsleistungen steigen dramatisch (im Mittel um 13,5% absolut bzw. 19,1% relativ) an, wenn die getesteten Äußerungen länger als 10 Sekunden andauern. Computer brauchen also wesentlich länger zu Identifizierung einer Sprache, als der Mensch.

3. Je feiner die Modellierung der akustischen Ereignisse wird (1-Zustand-HMM, broad phonemes, fine phonemes), desto besser scheint die LID zu funktionieren. Wie [40] und [21] schon feststellten (vgl. oben), wird die schlechtere Klassifikation durch den Mehrgehalt an Informationen, die bei feinerer Modellierung zur Identifikation zur Verfügung steht, mehr als kompensiert. Insbesondere scheint ein einzelnes Modell die Komplexität einer Sprache nicht ausreichend zu modellieren bzw. eine Unterscheidung verschiedener Sprachen zu ermöglichen.

4. Es kommen sehr unterschiedliche Systeme wie Li [19] und Zissmann et al. [40] mit völlig verschiedenen Informationsquellen zu nahezu denselben guten Leistungen, was den Schluß zuläßt, daß noch keine endgültige Entscheidung darüber gefällt werden kann, welche Ansätze für die LID am besten geeignet sind.

Kapitel 5

Experimente

Das Ziel dieser Diplomarbeit ist die Entwicklung eines sprachidentifizierenden Moduls für JANUS. Dieses Modul soll identifizieren, welche der möglichen Eingabesprachen Deutsch, Englisch oder Spanisch tatsächlich gesprochen wurde. In diesem nun folgenden Kapitel wird die Entwicklung des Moduls und die darauf durchgeführten Experimente beschrieben. Im ersten Abschnitt wird die multilinguale Datenbasis SST **Spontaneous Scheduling Task**, auf deren Basis die Experimente durchgeführt wurden, beschrieben. Im zweiten Abschnitt werden zunächst die Identifizierungsleistungen der aus der Literatur bekannten Ansätze auf der Datenbasis SST festgestellt. Dabei werden im dritten Abschnitt neue Ideen zur Auswertung der LID-Leistung vorgestellt. Der vierte Abschnitt ist der Analyse der LID-Leistung im zeitlichen Verlauf gewidmet. Die ersten vier Abschnitte machen den großen Einfluß des Aufnahmekanals auf die LID-Leistung deutlich. Dies wird im fünften Abschnitt detailliert veranschaulicht. Anschließend wird im sechsten Abschnitt anhand von kanalunabhängigen Experimente versucht, diesen Einfluß zu eliminieren. Im siebten, achten und neunten Abschnitt werden die neu entwickelten sprachidentifizierenden Systeme mit dem in Abschnitt 2 beschriebenen Referenzsystem verglichen. In diesen Abschnitten werden die Methoden an den beiden Sprachen Deutsch und Englisch dargestellt, denn in diesen beiden Sprachen wurden für diese Arbeit „Cross-channel"-Daten erhoben. Für Spanisch standen jedoch keine „Cross-channel"-Daten zur Verfügung. Daher werden im letzten Abschnitt die Experimente mit den spanischen Daten gesondert behandelt. Dieser Abschnitt widmet sich der Beschreibung des 3-Sprachen-Problems, d.h. der Unterscheidung von Deutsch, Englisch und Spanisch.

5.1 Die multilinguale Datenbasis SST

An der Universität Karlsruhe wurde in Kooperation mit der Carnegie Mellon University (Pittsburgh USA) im Oktober 1993 begonnen, eine multilinguale Datenbasis zu erstellen. Diese Datenbasis trägt den Namen **Spontaneous Scheduling Task (SST)** und enthält spontansprachliche Terminabsprachedialoge zwischen zwei Partnern. Die beiden Gesprächspartner werden aufgefordert, während einer sogenannten Dialogsitzung einen

Termin für ein fiktives Treffen zu vereinbaren. Zu diesem Zweck erhalten beide ein Kalenderformblatt, auf dem neben den Tagen eines gesamten Monats auch Feiertage und feste Termine eingetragen sind. Auf diese Weise soll der Gebrauch von typischen Begriffen rund um das Terminabspracheszenario angeregt werden. Die Formblätter sind für die Gesprächspartner verschieden, um Terminkollisionen hervorzurufen und dadurch den Dialog möglichst realitätsnah zu gestalten. Die Partner sitzen während des Dialogs gemeinsam in einem geschlossenen Raum, um zu viele Nebengeräusche wie Telefonklingeln, Papierrascheln, Hintergrundgespräche usw., die durch die Büroumgebung entstehen könnten, zu vermeiden. Während des Dialogs wird das Sprachrecht durch ein Knopfdruckverfahren geregelt. Nur derjenige Sprecher, der einen Knopf betätigt hat, darf sprechen. Die Sprache der Dialogpartner wird auf separaten Kanälen aufgenommen. Ein Dialog enthält in der Regel zwischen 10 bis 15 Äußerungen.

Englische SST		
	Dialoge	Worte
aufgenommen	1984	505 K
transkribiert	1826	460 K
Deutsche SST		
	Dialoge	Worte
aufgenommen	734	158 K
transkribiert	534	115 K
Spanische SST		
	Dialoge	Worte
aufgenommen	340	79 K
transkribiert	256	70 K

Tabelle 5.1: Aktueller Stand der Datenbasis SST

Die Daten werden sowohl in Karlsruhe (KA) als auch an der Carnegie Mellon University in Pittsburgh, USA (CMU) erhoben. In Karlsruhe werden deutsche Dialoge von deutschsprachigen Sprechern gesammelt, in Pittsburgh amerikanische und spanische Dialoge, jeweils von muttersprachlichen Sprachspendern. Das Szenario, die Kalender und die Datensammelprozedur verlaufen dabei nach festgelegten Modalitäten und werden in Protokollen festgehalten, um eine Vergleichbarkeit der gesammelten Daten an allen Orten zu gewährleisten. Die Tabelle 5.1 zeigt den derzeitigen Stand (Februar 1995) der gesammelten Datenbasis [35]. Mit der Sammlung spanischer Sprachdaten wurde erst kürzlich begonnen, daher stehen in Spanisch noch nicht soviele Trainingsdaten zur Verfügung.

Inzwischen haben viele Einrichtungen in Europa, den Vereinigten Staaten und Asien die SST-Datenbasis in verschiedenen Projekten eingesetzt. Derzeit werden Dialoge in koreanischer Sprache gesammelt, und die Aufnahme japanischer Dialoge beginnt demnächst. Zwar verfügt man mit der Datenbasis SST in nächster Zeit nur über 5 Sprachen gegenüber 11 Sprachen der multilingualen Datenbasis OGI, aber bei den SST Sprachdaten handelt es

sich um spontan gesprochene Mensch-zu-Mensch Äußerungen, nicht um einseitig geführte Telefonanrufe wie bei der OGI-Datenbasis. Zur Evaluation von LID-Systemen, die die Identifizierung von Sprache in spontansprachlichen Dialogen zwischen Menschen zum Ziel hat, könnte die SST-Datenbasis den OGI-Task in Zukunft ablösen.

Zu Beginn der Diplomarbeit standen die in Tabelle 5.1 genannten Daten noch nicht im vollen Umfang zur Verfügung. Die Tabelle 5.2 zeigt den Datenstand für die Diplomarbeit. Die verfügbaren Daten wurden in eine Trainings- und eine Testmenge unterteilt.

Task	Dialoge	Äußerungen	Worte
Trainingsdaten			
ESST	117	1205	38809
GSST	192	1985	45034
SSST	75	1270	61382
Testdaten			
ESST	20	173	4731
GSST	18	142	4107
SSST	13	86	3740

Tabelle 5.2: Trainings- und Testdatenmaterial

Die Länge der gesprochenen Äußerungen variiert zwischen den Sprachen sehr stark. Die Grafik 5.1 zeigt die Länge der Testsätze in Sekunden. Man sieht, daß die spanischen Testäußerungen eher länger sind, die englischen Testäußerungen sind dagegen sehr kurz.

5.2 Der traditionelle Ansatz

Zu Beginn der Experimente sollte ein Referenzsystem geschaffen werden. Dazu wurden bereits implementierte veröffentlichte Ansätze mit dem JANUS-System nachgebildet. Das erste Referenzsystem ist der parallele akustisch-phonetische Ansatz, der bereits in Kapitel 4.2.1 ausführlich beschrieben wird. Die akustische Modellierung geschieht hier auf Phonembasis wie etwa in [17], [36] und [10]. Dieser Ansatz soll in Zukunft mit PohneLM bezeichnet werden. Beim zweiten Referenzsystem wird der parallele phonologische Ansatz mit integriert berechneter sprachlicher Modellierung, wie etwa bei Lamel und Gauvain [17] beschrieben, nachgebildet. Dieses System wird im folgenden mit PmitLM bezeichnet.

Es wurde für die Sprachen Englisch und Deutsch je ein Phonemerkenner konzipiert. Zur akustischen Modellierung wurden beim englischen Erkenner 54 kontextunabhängige Phoneme mit SCHMMs modelliert, beim deutschen Erkenner 47 Phoneme auf die gleiche Art. Beim englischen Phonemerkenner befinden sich unter den 54 Phonemen 10 spezielle

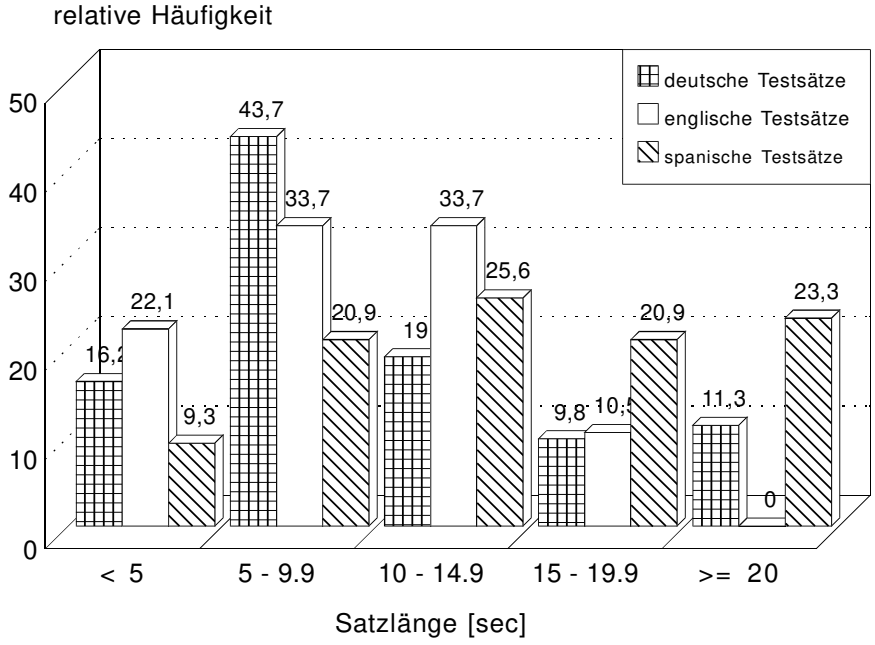

Abbildung 5.1: Länge der Testsätze in Sekunden

Geräuschmodelle, beim deutschen Erkenner 6 Geräuschmodelle. Die akustische Modellierung von sprachlichen und nichtsprachlichen Geräuschen hatte sich für spontan gesprochene Sprache als sehr nützlich erwiesen [32]. Geräusche, die durch den menschlichen Artikulationsapparat erzeugt werden, sind in der SST-Datenbasis von Sprache zu Sprache unterschiedlich. Beispielsweise wird im Englischen sehr häufig die Häsitation `uh` verwendet, im Deutschen dagegen eher das `äh` oder `ah`. Dies weist darauf hin, daß solche nonverbalen Produktionen durchaus sprachentypisch sein könnten. Daher sollen die Geräuschmodelle zur Identifizierung von Sprachen mit in den Phonemsatz aufgenommen werden.

Die Phonembigramme, die beim phonologischen Ansatz eingesetzt werden, wurden auf dem vorhandenen Trainingsmaterial (Tabelle 5.2) mit der *absolute discounting*-Methode geschätzt. Die Phonemerkennungsleistung der beiden Systeme PohneLM und PmitLM in Tabelle 5.3 ist vergleichbar mit der von Muthusamy et al. auf der OGI-Datenbasis ermittelten Leistung [21]. Lamel und Gauvain [17] erreichten dagegen Phonemerkennungsleistungen von 78,7% und 73,4% allerdings für gelesene Sprachdaten, nicht für spontane Daten.

Da die Geräusche keine semantische Bedeutung tragen, werden sie nicht als übersetzungsrelevante Teile angesehen. Daher ist es nicht sinnvoll, sie zur Weiterverarbeitung in den Hypothesen zu belassen. Ziel der Geräuschmodellierung ist es, die vorhandenen Geräusche zu erkennen, um sie anschließend eliminieren zu können. Aus diesem Grund wird die Erkennungsleistung nicht von den tatsächlich erkannten Satzhypothesen berech-

Testsätze	PohneLM	PmitLM
deutsch (142)	38,6%	48,8%
englisch (173)	39,3%	46,2%
spanisch (86)	32,2%	46,4%
Gesamt (401)	37,5%	47,2%

Tabelle 5.3: Phonemerkennungsleistung

net, sondern von den Sätzen, aus denen vorher allen Geräusche entfernt wurden. Entsprechend werden zur Berechnung der Phonemerkennungsleistung auch alle Geräuschvorkommen aus den Transkriptionen der Referenzsätze eliminiert. Damit ergibt sich die folgende Phonemerkennungsleistung in Tabelle 5.4

Testsätze	PohneLM	PmitLM
deutsch (142)	41,2%	49,6%
englisch (173)	42,6%	48,3%
spanisch (86)	33,9%	46,9%
Gesamt (401)	40,2%	48,5%

Tabelle 5.4: Geräuschbereinigte Phonemerkennungsleistung

Wie bereits in Kapitel 4 ausführlich beschrieben, basiert das Prinzip der Spracheniden-tifizierung mit einer parallelen Architektur darauf, daß ein spracherkennendes System stets in der Sprache, in der es trainiert wurde, bessere Bewertungen der Hypothesen erzielt als in einer fremden Sprache. Je nach Ansatz handelt es sich bei den Bewertungen der Hypothesen um Wahrscheinlichkeiten oder um Distanzwerte, die im folgenden als *Score* bezeichnet werden.

Zur Identifizierung der Sprache wird eine Testäußerung beiden Erkennern, sowohl für die deutsche als auch für die englische Sprache, vorgelegt. Für ein- und dieselbe Testäußerung erhält man somit zwei Bewertungen, jeweils eine vom deutschen Erkenner (GSST-Erkenner) und eine vom englischen Erkenner (ESST-Erkenner). Die Abbildung 5.2 zeigt die Bewertungen aller Testäußerungen für den akustisch-phonetischen Ansatz PohneLM, Abbildung 5.3 entsprechend für den phonologischen Ansatz PmitLM. Die Abbildungen zeigen auf der Ordinate die Bewertungen des ESST-Erkenners, auf der Abszisse die des GSST-Erkenners. Die bewerteten Testäußerungen sind gemäß der Sprache, in der sie gesprochen wurden, mit verschiedenen Markierungen gekennzeichnet. Wie man erkennt, sind die Sprachen Englisch und Deutsch auf der Basis beider Ansätze sehr gut separierbar. Da jedoch die Unterschiede zwischen den beiden Ansätzen grafisch nicht so deutlich erkennbar sind, und andererseits aus Gründen der Übersicht nicht zu viele Abbildungen gezeigt

werden können, sind im folgenden nur noch die Grafiken für den phonologischen Ansatz abgebildet. Im nächsten Abschnitt wird erläutert, wie die Identifizierungsleistung berechnet werden kann.

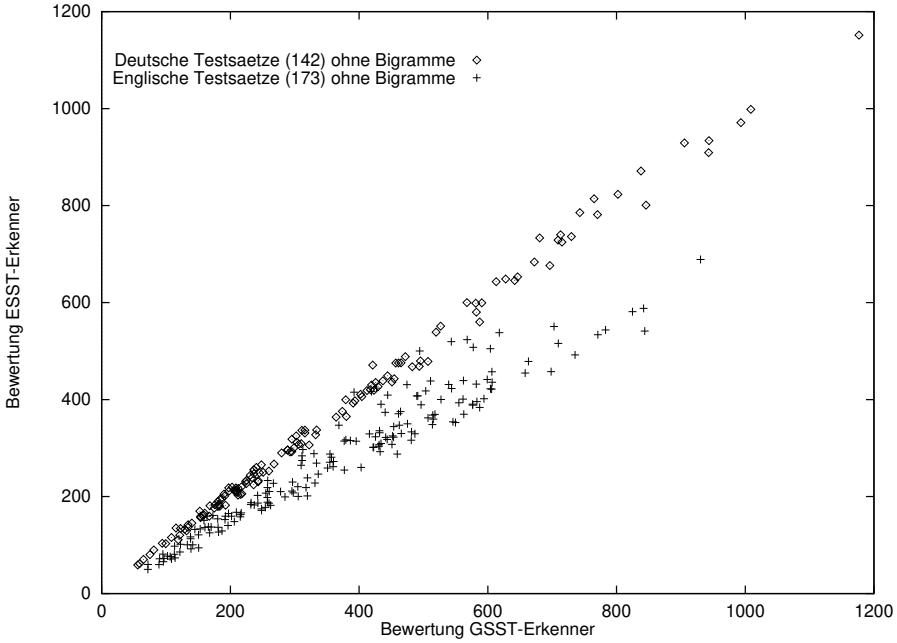

Abbildung 5.2: LID mit akustisch-phonetischem Ansatz

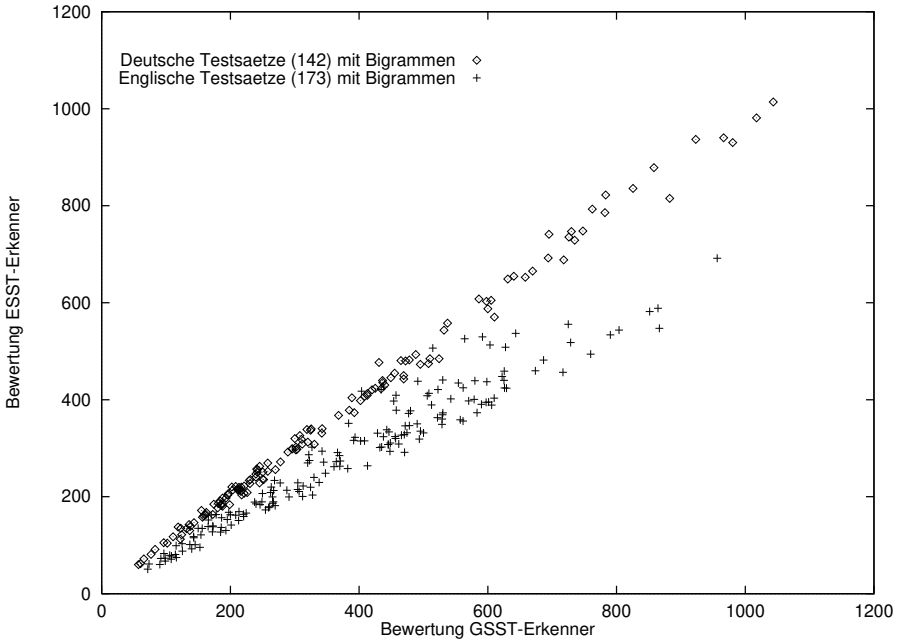

Abbildung 5.3: LID mit phonologischem Ansatz

5.3 Die Berechnung der LID-Leistung

Die Leistung eines sprachidentifizierenden Systems wird durch den Vergleich der Hypothesenscores der beteiligten Erkenner ermittelt. Im allgemeinen berechnet man den prozentualen Anteil der korrekt identifizierten Sätze für jede Sprache getrennt. Häufig wird zusätzlich ein über die Testäußerungen aller Sprachen gewichteter Mittelwert in Prozent angegeben. Bei der parallelen Architektur, d.h. der Verwendung eigenständiger sprachspezifischer Erkenner, können die Bewertungen der Hypothesen auf verschiedenen Grundlagen basieren. Beim JANUS-System wird eine Hypothese nicht anhand einer Wahrscheinlichkeit bewertet, sondern anhand eines Scores, der durch Akkumulation von Distanzen über den gesamten Viterbipfad zustandekommt. Anders als bei der Verwendung von Wahrscheinlichkeiten ist dieser Score u.a. von dem zugrundeliegenden Codebook, auf dem die Distanzen berechnet werden, abhängig. Aus diesem Grund sind die Hypothesenbewertungen des deutschen Erkenners nicht absolut vergleichbar mit denen des englischen Erkenners. Kleinere Distanzwerte sind somit nicht notwendigerweise gleichbedeutend mit einer besseren Bewertung. Daher kann zur Identifizierung der Sprachen nicht einfach die Winkelhalbierende als Trenngerade verwendet werden[1], sondern es sollte eine Normierung vorgenommen werden. Zwei Möglichkeiten bieten sich zur Normierung der Bewertungen an: einerseits eine Transformation des Raumes durch die Normierung der von den Erkennern berechneten Hypothesenbewertungen, andererseits die Berechnung einer von der Winkelhalbierenden abweichenden Trenngeraden.

5.3.1 Normierung durch Subtraktion des Mittelwertes

Falls die Bewertungen zweier Erkenner nicht nur zufällig voneinander verschieden sind, ist die Differenz der Mittelwerte über alle Bewertungen eines jeden Erkenners ein recht gutes Maß für den Unterschied. Für jeden Erkenner wird daher durch die Bildung des Mittelwertes über die Scores aller von diesem Erkenner bewerteten Testsätze der erkennerspezifische Mittelwert berechnet. Dieser erkennerspezifische Mittelwert wird vom eigentlichen Score subtrahiert. Die Differenz bildet die erkennerbereinigte, normierte Hypothesenbewertung. Anschließend wird zur Ermittlung der LID-Leistung die *Maximum Likelihood*-Regel angewendet. Dieses Verfahren wurde von Zissmann vorgeschlagen (vgl. [40]). Die Tabelle 5.5 zeigt die berechneten Mittelwerte der Bewertungen aller Testsätze.

Die deutschen Testsätze sind, wie in Abbildung 5.1 gezeigt, im Mittel wesentlich länger als die englischen. Daher ist der Mittelwert der Hypothesenbewertungen bei den deutschen Sätzen viel größer. Während aber bei den deutschen Sätzen die Scores beider Erkenner in ihrer Größenordung nahe beieinander liegen, fällt bei den englischen Testäußerungen auf, daß die Bewertungen beider Erkenner sehr weit auseinandergehen. Auf dieses Phänomen wird im Abschnitt „Kanalabhängigkeit“ noch einmal eingegangen. Die Scoredifferenz zwischen dem englischen und dem deutschen Erkenner auf den englischen Hypothesen wird

[1] Die Verwendung der Winkelhalbierenden soll in Anlehnung an das für den Vergleich von Wahrscheinlichkeiten übliche Verfahren als *Maximum Likelihood*-Regel bezeichnet werden

System	GSST	ESST
deutsche Sätze (142)		
PohneLM	360,98	366,28
PmitLM	364,80	364,76
englische Sätze (173)		
PohneLM	374,39	283,95
PmitLM	385,85	286,03
alle Sätze (315)		
PohneLM	368,35	321,07
PmitLM	376,40	321,39

Tabelle 5.5: Erkennerabhängige Mittelwerte der Hypothesenscores

mit zunehmender Satzlänge größer. Dies ist auch der Grund, weshalb diese Art der Normierung in dem vorliegenden Fall zu ungünstigen bzw. verfälschten Resultaten führt. Die Subtraktion eines globalen, unabhängig von der Satzlänge errechneten Mittelwertes führt zu einer noch akzeptablen Trennung bei langen Testäußerungen, kurze Testsätze werden jedoch nicht gut getrennt.

Die Normierung wurde für den akustisch-phonetischen und den phonologischen Ansatz durchgeführt. Die Abbildung 5.4 zeigt die Situation nach der Normierung für den phonologischen Ansatz. Wie man erkennt, wird durch die Lage der Punkte in der Ebene die korrekte Klassifikation deutscher Testäußerungen stark begünstigt. Aus der Tabelle 5.6 sind die LID-Leistungen des Systems auf der Basis der Normierungsart *Subtraktion des Mittelwertes* zu entnehmen. Die aufgezeigte fehlerfreie Identifizierung deutscher Sätze zeugt nicht von einer hohen Leistungsfähigkeit des Systems. Vielmehr wird die Asymmetrie der Trennung durch das Normierungsverfahren ungenügend kompensiert. Daher gehen die 100% LID-Leistung bei deutschen Sätzen zu Lasten von 67% bei den englischen Sätzen. Von einem guten Normierungsverfahren wird eine symmetrische Trennung erwartet. Im übernächsten Abschnitt wird gezeigt, daß die Berechnung der Trenngeraden durch ein neuronales Netz diese Forderung erfüllt.

System	Deutsch	Englisch	Gesamt
PohneLM	100%	67,1%	81,9%
PmitLM	98,6%	66,5%	81,0%

Tabelle 5.6: LID-Leistungen bei der Normierung durch Subtraktion des Mittelwertes

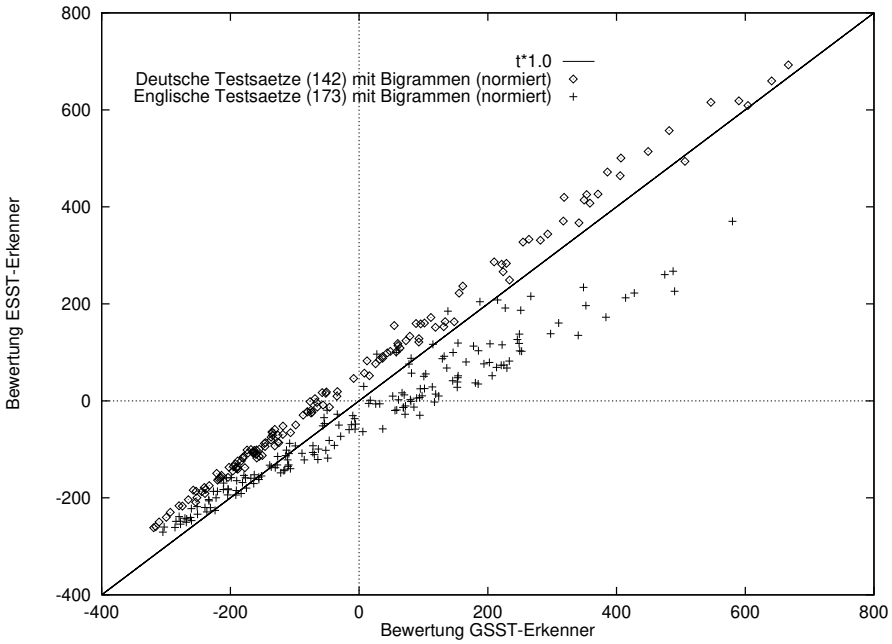

Abbildung 5.4: Normierung durch Subtraktion der Mittelwerte

5.3.2 Normierung durch Kalibrieren

Eine weitere Möglichkeit der Normierung, die den Nachteil der Abhängigkeit von der Dauer einer Testäußerung vermeidet, ist die *Kalibrierung* des Systems durch Ermittlung des mittleren Scores pro bewertetem Sprachframe über alle Trainingssätze. Man berechnet damit ein Maß, zukünftig als *Scoreperframe* bezeichnet, das für jeden Erkenner angibt, wie groß im Mittel die Bewertung eines Testframes ist. Dividiert man nun die Gesamtbewertung der Hypothese eines Erkenners durch dessen erkennerspezifischen *Scoreperframe*, erreicht man eine von der Länge der bewerteten Hypothese unabhängige Normierung. In der durch die Bewertungen aufgespannten Ebene bedeutet dies eine lineare Verschiebung der Punkte. Anschließend wendet man die *Maximum Likelihood*-Regel an. Diese Normierungsart führt zu den in der Grafik 5.5 veranschaulichten Ergebnissen, dargestellt am phonologischen Ansatz. Wie man daraus und aus Tabelle 5.7 sieht, führt diese Art der Normierung zu signifikant besseren Ergebnissen als die Normierung durch Subtraktion. Dies liegt vor allem an der besseren Unterscheidung kurzer Sätze.

System	Deutsch	Englisch	Gesamt
PohneLM	100%	88,5%	93,7%
PmitLM	100%	86,1%	92,4%

Tabelle 5.7: LID-Leistungen bei der Normierung durch Kalibrieren

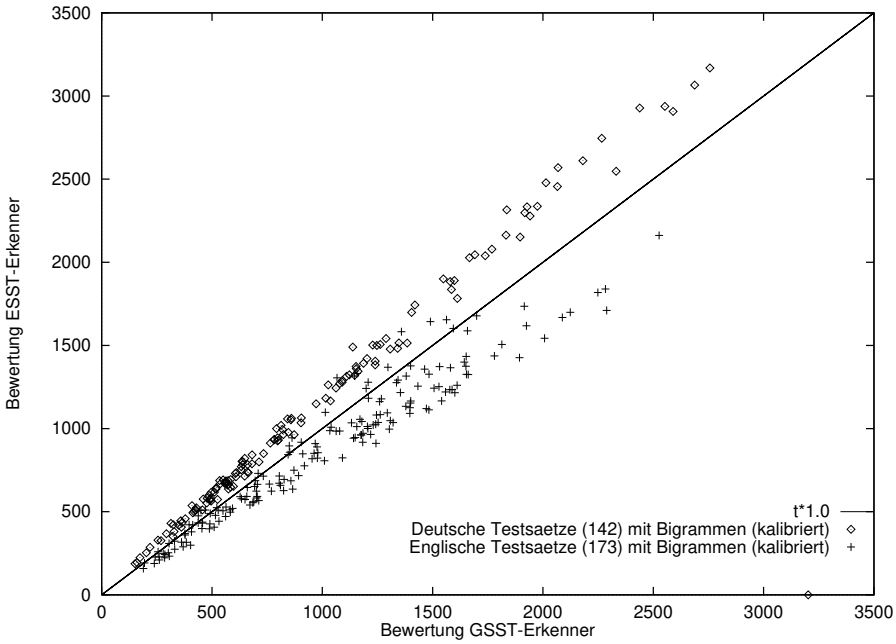

Abbildung 5.5: Normierung durch Kalibrieren

Beide Normierungsverfahren haben das Problem der asymmetrischen Trennung des Raumes und damit der Begünstigung der Identifizierung einer von zwei Sprachen nicht beheben können. Aus diesem Grund wurde nach einer weiteren Möglichkeit zur Normierung gesucht. Dabei wurde statt der Normierung des Raumes die Berechnung einer von der Winkelhalbierenden abweichenden Trenngeraden vorgenommen.

5.3.3 Berechnung einer Trenngerade mit Neuronalen Netzen

Die Berechnung einer Trenngeraden erfolgte mit Hilfe der künstlichen neuronalen Netze. Es wurde ein Perzeptron verwendet, das mit der Methode der fehlklassifizierten Lernstichprobe [24] trainiert wurde. Beim Lernen des Perzeptrons wurde der Nulldurchgang der Trenngeraden erzwungen. Diese Variante hatte sich nach mehreren Experimenten als günstig erwiesen. Sie scheint auch inhaltlich plausibel, da zwischen den Scores der beiden Erkenner eine eher lineare Beziehung besteht und gleichzeitig zum Zeitpunkt $t = 0$ die Scores beider Erkenner 0 betragen.

Zum Training des Perzeptrons wurden die ersten 28 Testsätze jeder Sprache als Muster vorgegeben. Die Angaben zu den Identifizierungsleistungen beziehen sich dennoch auf die gesamte Menge aller Testsätze, inklusiver derer, die zum Training des Perzeptrons verwendet wurden. Diese Vorgehensweise wurde aufgrund mangelnden Testmaterials beschritten. Eine Absplittung von 28 Testsätzen als Crossvalidierungsmenge war vor allem für die kanalunabhängigen Experimente in den nachfolgenden Kapiteln nicht vertretbar. Der „beschönigende“ Effekt dieser Vorgehensweise ist allerdings gering, wie Testmessungen

ergaben.

Das Ergebnis der Normierungsart *Berechnung der Trenngeraden durch ein Perzeptron* für den phonologischen Ansatz zeigt die Abbildung 5.6.

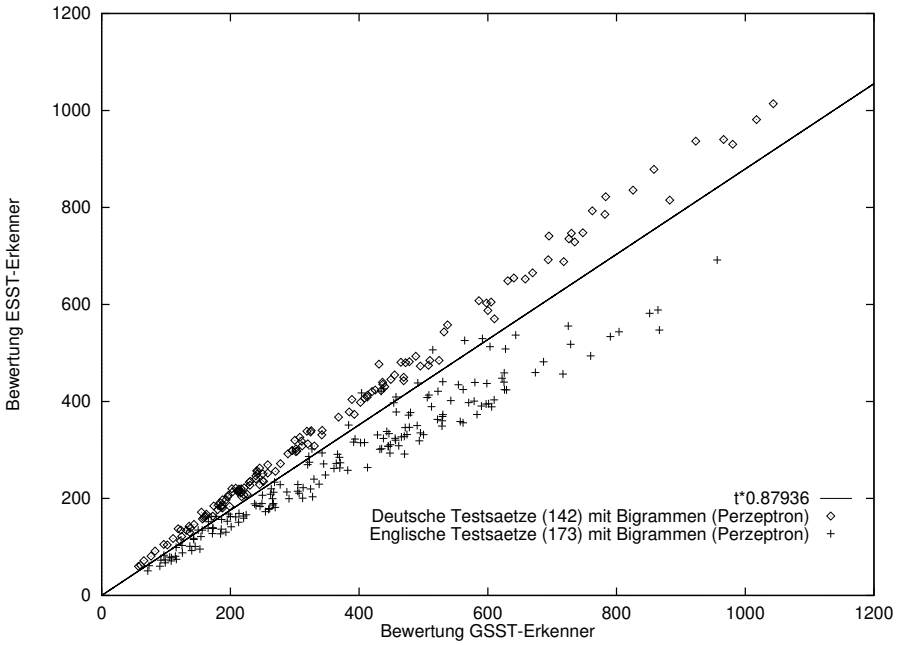

Abbildung 5.6: Berechnung der Trenngeraden durch Perzeptron

Die Tabelle 5.8 zeigt die LID-Leistungen für die mit einem Perzeptron errechnete Trenngerade.

System	Deutsch	Englisch	Gesamt
PohneLM	93,0%	96,5%	94,9%
PmitLM	100%	91,3%	95,2%

Tabelle 5.8: LID-Leistungen bei Berechnung der Trenngeraden durch Perzeptron

Die Berechnung einer Trenngeraden statt der Normierung des Raumes führt zu den mit Abstand besten Ergebnissen. Die Tabelle 5.9 zeigt die Leistungen bei der Unterscheidung der zwei Sprachen Englisch und Deutsch für die verschiedenen Normierungsarten verglichen damit, daß auf eine explizite Normierung verzichtet wird.

Allgemein gilt die Trennung zwischen deutscher und englischer Sprache als schwierige Aufgabe [9] [17] [39], während Spanisch gegen Englisch als gut trennbar eingeschätzt wird. Auf der OGI-Datenbasis wurde im Zweisprachentest von Zissmann [39] die Trennung von Englisch gegen Spanisch mit 83% angegeben, die Trennung von Englisch gegen Deutsch dagegen nur mit 67%. Lamel und Gauvain [17] stellten beim 10-Sprachentest fest,

Normierungsart	PohneLM	PmitLM
keine Normierung	86,3%	80,2%
Subtraktion	81,9%	81,0%
Kalibrierung	93,7%	92,4%
Perzeptron	94,9%	95,2%

Tabelle 5.9: Vergleich der verschiedenen Normierungsverfahren

daß Deutsch und Englisch am häufigsten miteinander verwechselt werden. In drei Studien wurden Leistungsangaben zur Identifizierung Englisch gegen Deutsch gemacht, leider konnte aber keine Untersuchung zur Trennung Deutsch gegen Englisch gefunden werden. Zissmann [39] kam mit ergodischen HMMs wie erwähnt auf 67%, Muthusamy [21] erreichte mit einem NN-Ansatz 77,7% und Barnard et al. erzielten 85,5% LID-Leistung [7] mit dem Polyphoneme-Ansatz. Die hier vorgelegten Ergebnisse für die Identifizierung der Sprachen Englisch und Deutsch liegen damit im Spitzenbereich im Vergleich mit veröffentlichten Ergebnissen. Wie bereits im Kapitel „Leistungsvergleich“ betont, ist aber ein Vergleich von Systemen, die Identifizierungsleistungen auf unterschiedlichen Daten messen, kaum möglich. Die Resultate sollten daher eher als Trend gewertet werden. Sie dienen hier vor allem als Vergleichspunkt für die in nachfolgenden Abschnitten dargestellten Systeme zur Identifizierung von Sprachen.

5.4 Zeitabhängige Identifizierungsleistung

Von einem guten sprachidentifizierenden System erwartet man nicht nur eine gute, sondern auch eine möglichst schnelle Identifizierung der Sprachen. Je länger die Sprachaufnahme der vorgelegten Testäußerung dauert, um so sicherer sollte ein System die gesprochene Sprache identifizieren können. Der Mensch ist in der Lage Sprachaufnahmen von 2 Sekunden Länge korrekt zu identifizieren. In dieser Geschwindigkeit ist er dem Computer eindeutig überlegen.

Es soll daher nun die Identifizierungsleistung des akustisch-phonetischen und des phonologischen Ansatzes in Abhängigkeit der Dauer des vorgelegten Sprachabschnittes analysiert werden. Die Abbildung 5.7 zeigt die Leistungen beider Systeme im Vergleich. Auf der Abszisse ist die Anzahl der dem System bereits vorgelegten Sprachframes aufgeführt, auf der Ordinate der prozentuale Anteil der richtig klassifizierten Testäußerungen. Die Kurven zeigen den prozentualen Anteil der korrekt klassifizierten Testsätze in Abhängigkeit der Dauer der vom sprachidentifizierenden System analysierten Sprachaufnahme. In der zeitabhängigen Berechnung der korrekt klassifizierten Äußerungen sind alle Sprachframes miteinbezogen. Das bedeutet, der Score eines jeden Sprachframes wird unabhängig davon, ob er sprachrelevante Informationen enthält oder nicht, zur Gesamtbewertung der Hypothese aufaddiert und damit zur Identifizierung der Sprache verwendet.

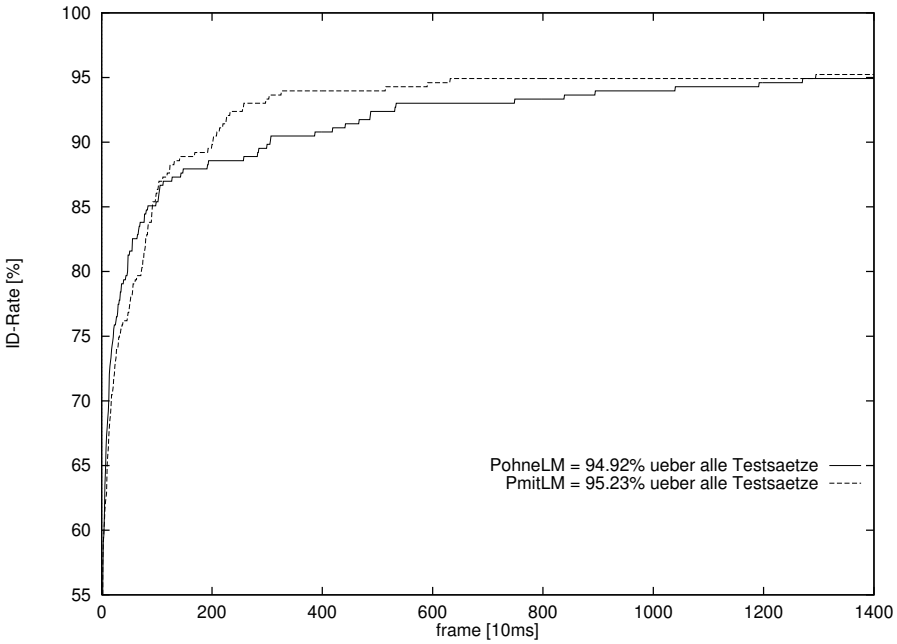

Abbildung 5.7: Zeitabhängige LID-Leistung mit und ohne Phonemgrammatik

Die Abbildung 5.7 und die Tabelle 5.9 zeigen, daß der phonologische Ansatz PmitLM mit einer LID-Leistung von 95,23% dem akustisch-phonetischen Ansatz PohneLM mit 94,92% in der absoluten Identifizierungsleistung leicht überlegen ist. Die absolute Leistungsdifferenz beträgt 0,31%, was einer relativen Fehlerreduktion von 6,1% entspricht. Die Reduktion des Fehlers durch den Einsatz der Phonembigramme ist für die Identifikation nicht so groß wie für die Erkennung von Sprache (etwa 25% laut Tabelle 5.3). Es scheint so, als seien die Unterschiede in der Sprache auf den reinen akustischen Merkmalen schon so groß, daß die zusätzliche linguistische Information nicht mehr viel zur Unterscheidung beiträgt. Anders ausgedrückt wird die Entscheidung für die eine oder andere Sprache fast ausschließlich auf akustischer Information gefällt. Dieses Phänomen wird im nächsten Abschnitt untersucht.

Die Grafik 5.7 zeigt aber auch, daß die Phonemgrammatik nach der ersten Sekunde zu einem schnelleren Ansteigen der Identifizierungsleistung führt, d.h. das zusätzliche linguistische Wissen verhilft dem System vor allem zur *schnelleren* Identifizierung der Sprache. Während mit Grammatik die nahezu endgültige Leistung schon nach 3,3 Sekunden Sprache erreicht wird, kommt es ohne Grammatik erst nach 6 Sekunden zu einer allmählichen Sättigung der Identifizierungskurve. Linguistisches Wissen in Form von Phonembigrammen kann somit zur besseren Identifizierung kurzer Sprachabschnitte beitragen. Diese Ergebnisse stehen im Einklang mit den Beobachtungen von Lamel und Gauvain [17] bei Untersuchungen an französischen und englischen Sprachdaten.

Bei den bisher gezeigten Ergebnissen wurde die Gesamthypothesenbewertung zugrundegelegt, die auf jeder Testäußerung ermittelt worden war. Diese Gesamthypothesenbewertung kommt durch Akkumulieren der für jeden Sprachframe ermittelten Scores entlang

des Viterbipfades zustande. Dabei wurde bisher der Score jedes Sprachframes akkumuliert, unabhängig davon, ob der jeweilige Sprachframe gesprochene Sprache enthält oder etwa Sprechpausen und längere Stilleperioden[2]. Wenn man die zeitabhängige Identifizierungsleistung eines Systems auf tatsächlich gesprochener Sprache messen möchte, müssen alle Sprachframes, die SILENCE enthalten, aus der Akkumulationsprozedur ausgeschlossen werden. Die LID-Leistung wird auf den daraus entstehenden SILENCE-*bereinigten* Hypothesenbewertung anschließend wie gewohnt berechnet.

Die nachfolgende Abbildung 5.8 zeigt den Verlauf der LID-Leistung über die Dauer der Testäußerungen für die deutschen Testsätze. Es wird die Identifizierungsleistung auf den SILENCE-bereinigten Hypothesenbewertungen (exklusive SILENCE) mit den bisherigen Gesamthypothesenbewertungen (inklusive SILENCE) für den phonologischen Ansatz PmitLM verglichen.

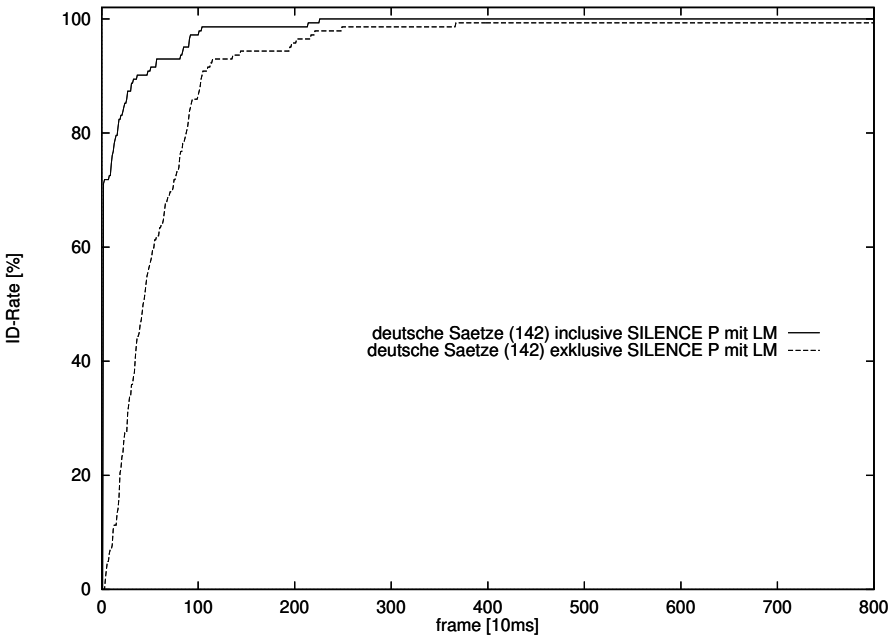

Abbildung 5.8: LID-Leistung inklusive und exklusive SILENCE auf deutschen Sätzen

Aus der Abbildung 5.8 ist ersichtlich, daß das spracherkennende System inklusive SILENCE schon nach den ersten Sprachframes eine Identifizierungsleistung von über 70% aufweist, während beim System exklusive SILENCE die Identifizierungskurve später und allmählich ansteigt. Daraus läßt sich folgern, daß die Unterscheidung der Sprachen nicht auf der Grundlage von tatsächlich gesprochener Sprache stattfindet. Die gute Identifizierungsleistung schon zu Beginn der Äußerung basiert vielmehr auf der Unterscheidung zwischen englischem und deutschem SILENCE. Ein SILENCE-Frames enthält unabhängig von der Sprache keine sprachrelevante Information, denn es handelt sich dabei schließlich um Stille oder Pause. Tatsächlich wird auch nicht englische von deutscher Stille unter-

[2]Für diese beiden Ereignisse wird im folgenden der Terminus *SILENCE* verwendet

schieden, sondern es werden die Hintergrund- bzw. Kanaleigenschaften der verschiedenen Aufnahmeorte erkannt.

Abbildung 5.9 zeigt die Verhältnisse für die englischen Testäußerungen ebenfalls für PmitLM. Im System exklusive SILENCE werden 6 englische Testäußerungen und eine deutsche Äußerung falsch klassifiziert, die vom System inklusive SILENCE korrekt klassifiziert worden waren. Es werden somit vom System Informationen aus den SILENCE-Frames entnommen, ohne die die Identifizierung nicht möglich ist.

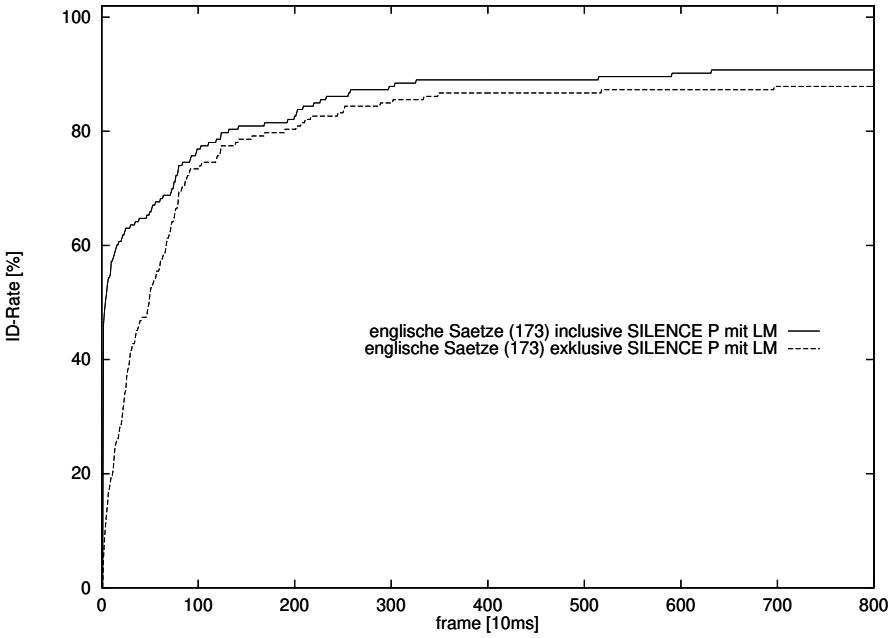

Abbildung 5.9: LID-Leistung inklusive und exklusive SILENCE auf englischen Sätzen

5.5 Das Problem der Kanalabhängigkeit

Die Beobachtungen aus den vorangegangenen Abschnitten sind alarmierend und führen zu einer Verfälschung der Ergebnisse. Sie machen deutlich, daß die bisher vorgestellten guten Identifizierungsleistungen beider Systeme nicht oder nicht nur von der sprachenbedingten Unterschiedlichkeit der Daten herrühren.

Um die Tragweite des Einflusses von Kanaleigenschaften auf die LID zu verdeutlichen, ist in der Abbildung 5.10 ein Ausschnitt des Identifizierungsgeschehens in den ersten 10-100ms abgebildet. Wie man sieht, „erkennt“ das System schon nach dem ersten Sprachframe zu fast 60% die richtige Sprache der Testäußerung, und dies, obwohl jede Äußerung mit SILENCE beginnt.

Die Tabelle 5.10 zeigt die im Mittel auf einen Sprachframe entfallenden Scores, die schon zum Kalibrieren des Systems eingesetzt wurden.

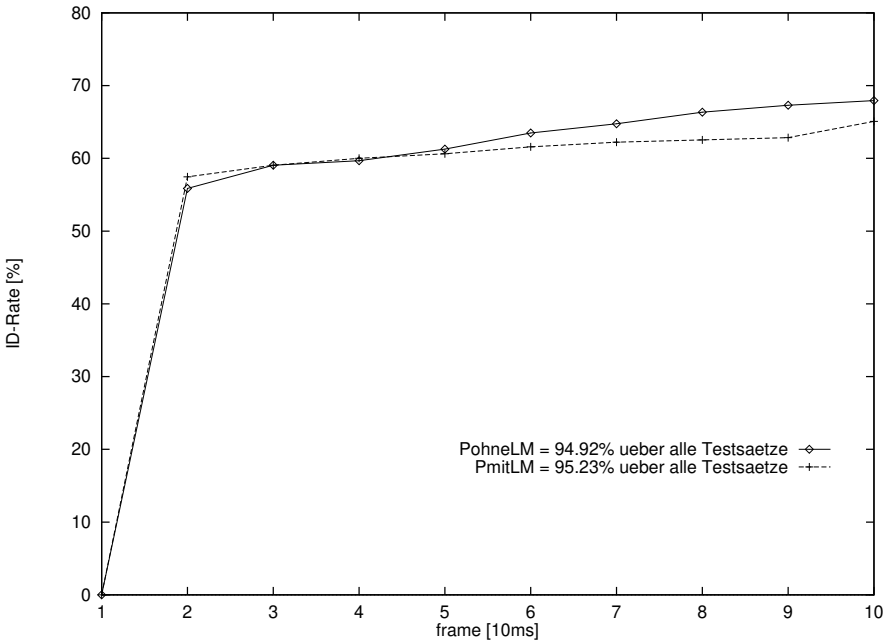

Abbildung 5.10: Erkennungsleistung in den ersten 10-100ms

Wie man beim zeilenweisen Vergleich in der Tabelle erkennt, sind die mittleren Scoresperframe auf den Testäußerungen der jeweils zum Erkenner gehörigen Sprache stets kleiner als auf der fremden Sprache. Dies zeigt was zu erwarten war, daß der Erkenner mit der Akustik, auf der er trainiert wurde, besser zurecht kommt. Auffallend ist aber der dramatische Anstieg des mittleren Scoreperframe des GSST-Erkenner auf englischen Testäußerungen. Während die Differenz der Mittelwerte zwischen deutschen und englischen Sätzen beim englischen Erkenner im Mittel nur 0,035 Punkte beträgt, bewertet der deutsche Erkenner die englischen Sätze um 0,06 Punkte schlechter.

Eine Überprüfung der akustischen Daten anhand des „signal-to-noise'" Quotienten ergab, daß die englischen Daten viel geräuschbehafteter sind als die deutschen. Obwohl bei der Sammlung der Daten sehr sorgfältig auf die Einhaltung der Aufnahmebestimmungen geachtet wurde, sind die Umstände, unter denen die Daten an den verschiedenen Orten (KA und CMU) aufgenommen wurden, nicht dieselben und können auch nur schwer beeinflußt werden. So fand die Aufnahme in Karlsruhe in einem separaten, abgetrennten Raum statt, in dem nur bedingt Umweltgeräusche auftraten. An der CMU mußten die Daten in einem Laboratorium aufgenommen werden, in dem Mitarbeiter und Studenten zeitgleich arbeiteten. Obwohl als Aufnahmemedium in beiden Fällen ein Nahsprechmikrophon derselben Marke verwendet wurde, waren die Verbindungskabel, die die Rechner verbinden, von unterschiedlicher Länge. Alle diese Umstände bedingen starke Unterschiede in der Qualität der aufgenommenen Daten.

Das Problem der Kanalabhängigkeit stellt sämtliche LID-Experimente, die auf Daten verschiedener Aufnahmeorte gemacht werden, in Frage. Und obwohl die in dieser Arbeit verwendeten Daten von stark kooperienden Instituten mit festgelegten gemeinsamen

System	GSST	ESST
deutsche Sätze (142)		
PohneLM	0,335	0,340
PmitLM	0,345	0,336
englische Sätze (173)		
PohneLM	0,399	0,302
PmitLM	0,411	0,304
alle Sätze (315)		
PohneLM	0,368	0,321
PmitLM	0,379	0,320

Tabelle 5.10: Erkennerabhängige mittlere Scoresperframe

Standards und gleicher Hardware aufgenommen wurden, sind die Einflüsse der Aufnahmeorte spürbar. In vielen Untersuchungen werden jedoch sogar Daten unterschiedlicher Domänen und verschiedener Institutionen verwendet. Trotzdem wird in der Literatur auf das Problem der Kanalidentifikation kaum eingegangen. In den folgenden Abschnitten dieser Arbeit werden nun Maßnahmen beschrieben, die zur Eliminierung der Kanaleinflüsse ergriffen wurden. Die Identifizierungsleistung wird auf tatsächlich gesprochenen Daten ohne den Einfluß unterschiedlicher Kanaleigenschaften gemessen werden. Ein Vergleich von kanalabhängigen und kanalunabhängigen Experimenten soll zeigen, wie hoch der Einfluß der Kanalidentifikation auf die LID-Leistung tatsächlich ist.

5.6 Kanalunabhängige Experimente

Um die Identifizierungsleistung des Systems ohne Effekte, die durch die Unterschiede in den Kanaleigenschaften entstehen, zu messen, wurde auf Cross-Channel Daten experimentiert. Dazu wurden 13 Dialoge in deutscher Sprache von Sprechern mit der Muttersprache Deutsch an der Carnegie-Mellon University unter den typischen CMU-Bedingungen aufgenommen (abgekürzt mit CMUger). Umgekehrt wurden 19 Dialoge in englischer Sprache von muttersprachlich englischen Sprechern in Karlsruhe aufgenommen (abgekürzt mit KA-eng). Dadurch ist nun ein Vergleich zwischen englischen und deutschen Sprachdaten der gleichen Kanaleigenschaften möglich.

Die Tabelle 5.11 zeigt die für diese Experimente aufgenommenen Daten. Die Datenaufnahmen wurden nicht transkribiert, somit können keine Angaben zur Anzahl gesprochener Worte gemacht werden.

Die Tabelle 5.12 zeigt den Einfluß des Kanals auf die mittleren Scoresperframe. Beim zeilenweisen Vergleich der Tabelle, d.h. dem Vergleich zwischen den Aufnahmeorten, sieht man, daß der GSST-Erkenner deutsche Sätze nahezu gleich bewertet, unabhängig davon, wo die Daten aufgenommen wurden. Die englischen Sätze werden wesentlich schlechter

Task	Dialoge	Äußerungen
KAeng	19	75
CMUger	13	82

Tabelle 5.11: Datenmaterial für die kanalunabhängigen Experimente

bewertet, wenn sie an der CMU aufgenommen wurden. Dieser Effekt, daß der Aufnahmeort CMU vom deutschen Erkenner schlechter bewertet wird, als der Ort Karlsruhe ist verständlich, da der deutsche Erkenner ausschließlich auf Karlsruher Aufnahmen trainiert wurde. Umso erstaunlicher, daß die deutschen Sätze aus der CMU nicht als „artfremd" schlechter bewertet werden. Im Gegensatz dazu ist beim ESST-Erkenner der Einfluß des Aufnahmeortes sowohl bei den deutschen als auch bei den englischen Testsätzen deutlich zu sehen.

Testäußerung	CMU-Daten	KA-Daten
GSST-Erkenner		
deutsche Sätze	0,330	0,335
englische Sätze	0,398	0,367
ESST-Erkenner		
deutsche Sätze	0,314	0,340
englische Sätze	0,302	0,334

Tabelle 5.12: Kanalabhängige mittlere Scoresperframe

Betrachtet man nun den vertikalen Vergleich, d.h. den Unterschied der Bewertungen in Abhängigkeit der gesprochenen Sprache, ist beim GSST-Erkenner eine große Differenz zwischen den Bewertungen zu sehen. Englische Sätze werden signifikant schlechter bewertet als deutsche Sätze, unabhängig davon, wo sie aufgenommen wurden. Für CMU-Aufnahmen wird die Diskrepanz größer. Beim ESST-Erkenner wird nun das Phänomen deutlich, daß deutsche Sätze, die an der CMU aufgenommen wurden, besser bewertet werden als englische Sätze, die in Karlsruhe aufgenommen wurden. Überspitzt formuliert heißt das, daß der ESST-Erkenner eher den Kanal als die Sprache, auf der er trainiert wurde, wiedererkennt.

Insgesamt scheint es für den deutschen Erkenner bezüglich der Bewertung der eigenen Sprache unerheblich, ob die Äußerungen in Karlsruhe oder an der CMU aufgenommen wurden. Für den englischen Erkenner jedoch sind in Karlsruhe aufgenommene Daten sprachlich schwer zu unterscheiden.

Die Abbildungen 5.11 zeigen den unterschiedlichen Einfluß des Kanals durch den Vergleich von Testsätzen der gleichen Sprache an den verschiedenen Aufnahmeorten Karlsruhe und CMU für das PohneLM System. Es läßt sich gut erkennen, daß bei den deutschen

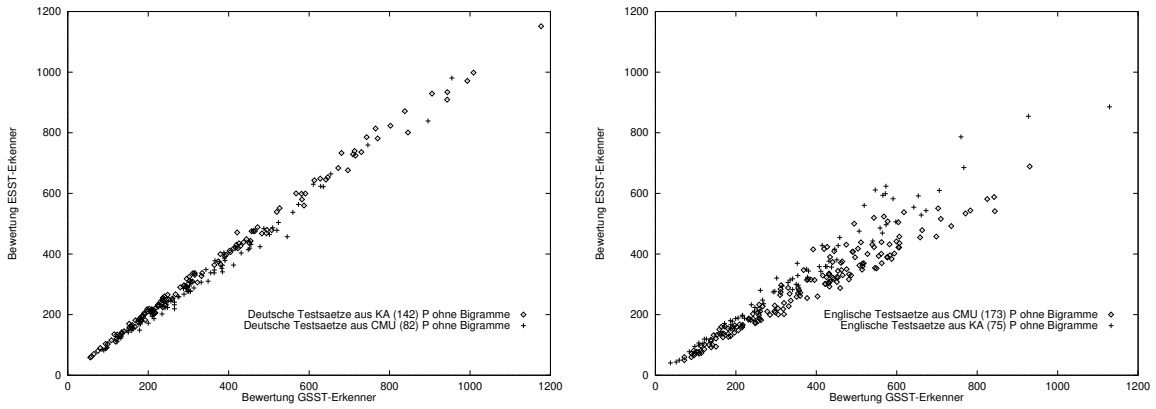

Abbildung 5.11: Kanaleinflüsse auf die Testdaten links für die deutschen und rechts für die englischen Testäußerungen

Sätzen keine Bewertungsunterschiede zwischen den Aufnahmeorten Karlsruhe und CMU bestehen. Die englischen Testäußerungen hingegen werden sichtlich unterschiedlich bewertet.

Nun ist die Frage, wie gut die LID-Leistung ist, wenn Daten gleicher Aufnahmeorte miteinander verglichen werden. Die Erkenner erhalten dazu die Daten der verschiedenen Sprachen an gleichen Orten, die somit mit gleichen Kanaleigenschaften aufgenommen wurden. Die Unterscheidung der Sprachen kann daher nicht mehr durch die Unterscheidung der Kanäle geschehen. Identifikationsleistungen, die das System auf diesen Daten bringt, sind somit nur auf die akustischen und sprachlichen Unterschiede der Sprachen zu beziehen. Eine Einschränkung ist allerdings die schon erwähnte Tatsache, daß die Erkenner auf verschiedener Datenqualität trainiert wurden. Der ESST-Erkenner wurde auf den Sprachaufnahmen der CMU trainiert, der GSST-Erkenner auf den Daten aus Karlsruhe. Der Einfluß, der dadurch entsteht, daß ein auf sauberen Daten trainierter Erkenner auf geräuschbehaftete Testdaten anders reagiert als ein auf geräuschbehafteten Daten trainierter Erkenner auf saubere Testdaten, kann leider nicht eliminiert werden, da für ein Training der Erkenner auf kanalfremdem Material nicht genügend Sprachdaten vorliegen. Andererseits gibt es keinen Grund zu der Annahme, daß dieser Effekt die Identifizierungsleistung beeinflußt. Diese Form der Experimente werden zukünftig als kanalunabhängige Messungen bezeichnet. Die Abbildung 5.12 zeigt die kanalunabhängigen LID-Leistungen exemplarisch für den Aufnahmeort CMU.

Tabelle 5.13 faßt die Leistungen von PmitLM und PohneLM für den Test auf Cross-Channel Daten zusammen. Die Identifizierungsleistung sinkt beim akustisch-phonetischen Ansatz von vorher 94,9% auf den kanalabhängigen Testdaten auf 88,8%. Beim phonologischen Ansatz reduziert sich die Leistung von 95,2% auf 88,8%. Durch die Eliminierung der Kanalidentifizierung verdoppelt sich somit die Fehlerrate. Insgesamt sind die Leistungen in der Identifizierung der Sprachen auf den in Karlsruhe aufgenommenen Daten signifikant schlechter als auf den CMU-Daten. Das liegt vor allem an den oben beschriebenen Problemen des ESST-Erkenners mit den Karlsruher Daten. Der Einsatz der Phonemgrammatik

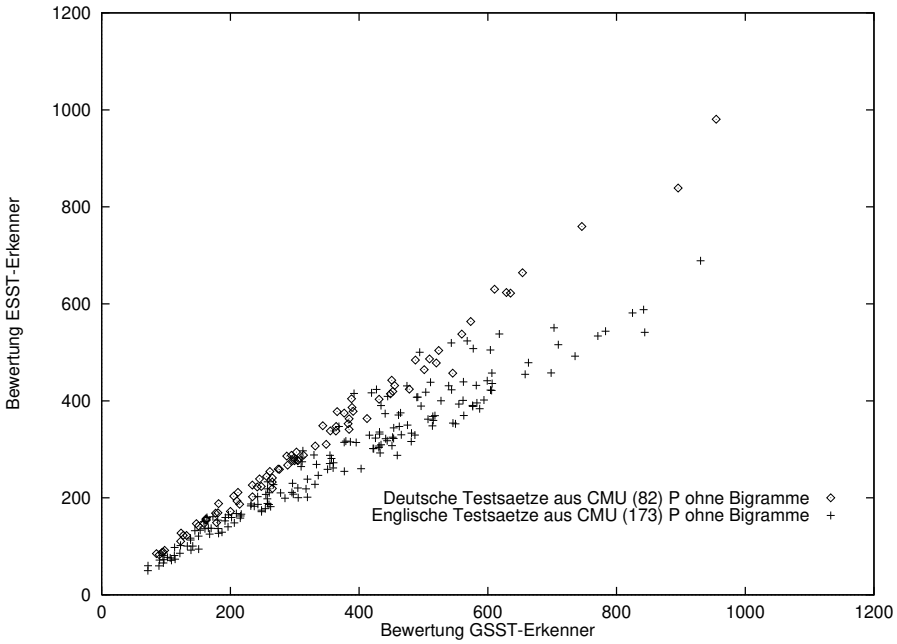

Abbildung 5.12: LID-Leistungen am Aufnahmeort CMU

führt bei den CMU-Daten zu einer Verbesserung der Leistung, bei den Karlsruher Daten hingegen zu einer Verschlechterung. In der Summe ist keine signifikante Veränderung durch die Grammatik zu verzeichnen. Für kanalabhängige Experimente betrug der Einfluß der Grammatik immerhin 0,3%, das entspricht einer relativen Fehlerreduktion von 6,8%.

Die LID-Leistung in Abhängigkeit der Zeit zeigt die Abbildung 5.13. Die jeweils zugehörigen Kurven für die deutschen Karlsruher und die englischen CMU-Daten wurden aus Gründen der Übersichtlichkeit weggelassen. Der Identifizierungsprozeß basiert hier nicht mehr auf Kanaleigenschaften. Man sieht nun deutlich, daß die LID-Leistungen erst langsam mit den gesehenen Sprachframes ansteigt, da die SILENCE-Frames keine zusätzlichen Informationen mehr bieten. Insgesamt sind die kanalunabhängigen LID-Leistungen nicht sehr befriedigend, und es sollen nun Systementwicklungen vorgestellt werden, die die Identifizierung der Sprachen verbessern.

System	Deutsch	Englisch	Gesamt
Aufnahmen in Karlsruhe			
PohneLM	93,7%	74,7%	87,1%
PmitLM	91,6%	76,0%	86,2%
Aufnahmen an der CMU			
PohneLM	90,2%	90,2%	90,2%
PmitLM	92,7%	90,2%	91,0%
alle Aufnahmen gemeinsam			
PohneLM	92,4%	85,5%	88,8%
PmitLM	92,0%	85,9%	88,8%

Tabelle 5.13: Kanalunabhängige LID-Leistung der Systeme PohneLM und PmitLM

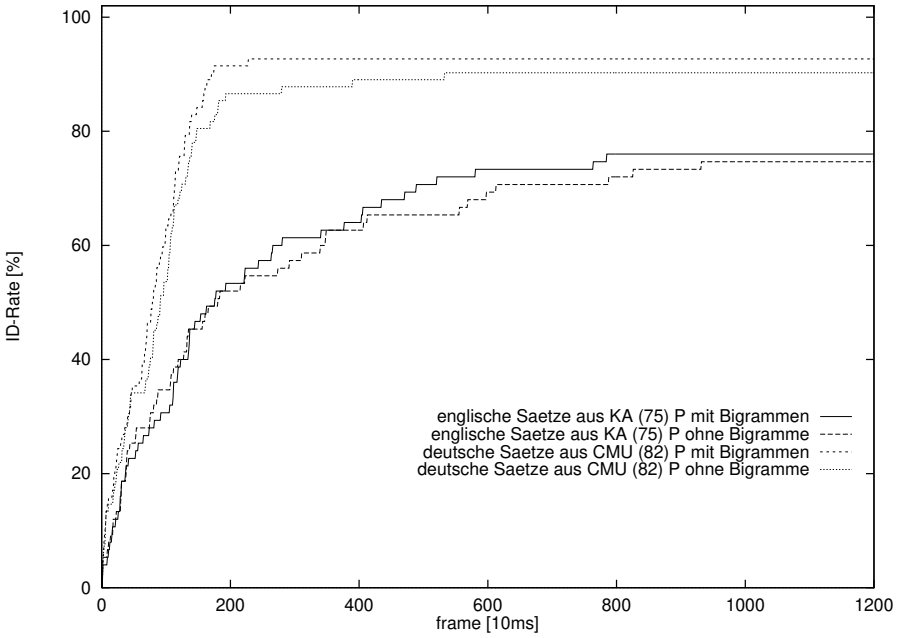

Abbildung 5.13: Kanalunabhängige LID-Leistung mit und ohne Grammatik

5.7 Integration von höherem Wissen

In den bisher dargestellten Systemansätzen wurden von den in Kapitel 2 beschriebenen Wissensquellen lediglich zwei ausgenutzt, nämlich die akustisch-phonetischen und die phonologischen Merkmale einer Sprache. Das Wissen über den Wortschatz und die grammatikalische Struktur einer Sprache wurde ignoriert. Wie bereits in Kapitel Forschungsansätze erwähnt, konnte bisher keine Arbeit im Bereich LID gefunden werden, die Wissensbasen wie Lexikon und Grammatik in den LID-Prozeß integrieren. Für Systeme, die ausschließlich Sprachen unterscheiden sollen, ist der mit der Lexikon- und Grammatikerstellung verknüpfte Aufwand auch außerordentlich hoch. Außerdem ist der mit der Erkennung auf Wortebenen verbundene Rechenaufwand viel größer als der auf Phonemebene. In dem hier vorliegenden Fall soll jedoch ein LID-System als „front-end" Modul für ein Sprache-zu-Sprache Übersetzungssystem konzipiert werden. Daher liegen Wissensbasen wie Lexikon und Grammatik bereits vor, weil sie für den Spracherkennungsprozeß benötigt werden. Außerdem kann die Information, für deren Berechnung auf Wortebene Rechenzeit „verschwendet" wird, für die anschließende Spracherkennung benutzt werden. Das Hinzufügen von Wissen in Form von Lexikon und Grammatik hat vor diesem Hintergrund keine Nachteile. Der Vorteil einer Integration des Lexikons liegt in der Möglichkeit, mit ausländischem Akzent gesprochene Sprache korrekt zu identifizieren. Eine Situation, in der ein Sprecher versucht Worte in einer ihm fremden Sprache auszusprechen und dabei seine muttersprachlichen Phoneme verwendet, ist in einem Dialog zwischen zwei Menschen verschiedener Muttersprachen gut vorstellbar. Die Verwendung einer Grammatik könnte dabei zusätzlich helfen. Im folgenden soll untersucht werden, ob das Hinzufügen dieser Wissensquellen eine signifikante Verbesserung der Identifizierungsleistung zur Folge hat. Zu diesem Zweck werden zwei neue Systemansätze eingeführt und anschließend mit den bisher beschriebenen Systemen PohneLM und PmitLM verglichen.

Die Integration eines Lexikons bedeutet, daß dem Spracherkenner Wissen über die Verknüpfung von Phonemen zu Worten eines Sprachwortschatzes hinzugefügt wird. Es handelt sich dann nicht mehr um eine Erkennung des Gesprochenen auf der Ebene von Phonemen, sondern auf der Ebene von Worten. Die akustische Modellierung basiert zwar nach wie vor auf den Subeinheiten der Phoneme, aber die Suche orientiert sich an ganzen Worten, wie bereits in Kapitel 3 beschrieben. Zwei Systeme CDohneLM und CDmitLM wurden verwendet:

- CDohneLM: Beim System CDohneLM werden kontextabhängige Phoneme verwendet. Dadurch werden zusätzlich zu den Systemen PohneLM und PmitLM Koartikulationseffekte modelliert, die zur Identifizierung benutzt werden können. Das System CDohneLM enthält ein Lexikon, dessen Einträge die Konkatenation der Phoneme aller verwendeteter Worte beschreibt. Der deutsche Erkenner hat einen Wortschatz von 2077 Worten, das englische Vokabular enthält 1073 Einträge.

- CDmitLM: Das System CDmitLM ist mit dem von CDohneLM identisch, enthält aber darüberhinaus grammtikalisches Wissen. Dieses Wissen liegt in Form einer Bigramm-Grammatik auf Wortebene vor.

Testsätze	CDohneLM	CDmitLM
deutsch (142)	23,0%	65,8%
englisch (173)	25,7%	65,2%
spanisch (86)	36,7%	63,6%
Gesamt (401)	27,1%	65,1%

Tabelle 5.14: Geräuschbereinigte Worterkennungsleistung

In einem ersten Schritt wurde die Worterkennungsleistung (analog zur Phonemerkennungsleistung) der beiden Systeme auf den Sprachen Deutsch, Englisch und Spanisch ermittelt. Die Ergebnisse sind in der Tabelle 5.14 für die geräuschbereinigten Hypothesen dargestellt. Für die Erkennung von Sprache liefert der Einsatz von Lexikon und Wortgrammatik eine entscheidende Leistungsverbesserung. Ob dies auch auf die Identifizierung von Sprachen zutrifft, ist eine der zentralen Fragen dieser Arbeit. Die Tabelle 5.15 zeigt den Leistungsvergleich aller vorgestellten Systeme PohneLM, PmitLM, CDohneLM und CDmitLM in den kanalunabhängigen Experimenten.

System	Deutsch	Englisch	Gesamt
Aufnahmen in Karlsruhe (d142 vs e75)			
PohneLM	93,7%	74,7%	87,1%
PmitLM	91,6%	76,0%	86,2%
CDohneLM	91,6%	78,7%	87,1%
CDmitLM	93,7%	81,3%	89,4%
Aufnahmen an der CMU (d82 vs e173)			
PohneLM	90,2%	90,2%	90,2%
PmitLM	92,7%	90,2%	91,0%
CDohneLM	97,6%	88,5%	91,4%
CDmitLM	92,7%	93,6%	93,3%

Tabelle 5.15: Kanalunabhängige LID-Leistung aller Systeme im Vergleich

Der Leistungsvergleich der Systeme zeigt, daß für die CMU-Daten die Integration sowohl des Lexikons als auch der Wortgrammatik die Identifizierung von Sprachen entscheidend verbessert. Für die Karlsruher Daten ist es das System CDmitLM, durch welches eine große Leistungssteigerung erzielt wird. Tabelle 5.16 zeigt die Verbesserungen für die

kanalabhängigen Experimente. Auch hier brachte die Integration weiterer Wissensquellen einen wichtigen Leistungszuwachs. Die Tabelle 5.17 faßt die Reduktion des Identifizierungsfehlers für die kanalabhängigen und die kanalunabhängigen Experimente zusammen. Die Fehlerreduktion ist in Prozent angegeben und bezieht sich auf das Referenzsystem PohneLM am jeweiligen Aufnahmeort.

System	Deutsch	Englisch	Gesamt
Kanalabhängiger Vergleich (d142 vs e173 - 28)			
PohneLM	93,0%	96,5%	94,9%
PmitLM	100%	91,3%	95,3%
CDohneLM	100%	95,1%	97,3%
CDmitLM	100%	96,6%	98,1%

Tabelle 5.16: Kanalabhängige LID-Leistung aller Systeme im Vergleich

System	relative Fehlerreduktion
Aufnahmen in Karlsruhe (d142 vs e75)	
PohneLM	Referenzpunkt
PmitLM	-7,1%
CDohneLM	0,0%
CDmitLM	17,9%
Aufnahmen an der CMU (d82 vs e173)	
PohneLM	Referenzpunkt
PmitLM	8,2%
CDohneLM	12,2%
CDmitLM	31,9%
Kanalabhängiger Vergleich (d142 vs e173)	
PohneLM	Referenzpunkt
PmitLM	7,8%
CDohneLM	47,0%
CDmitLM	62,6%

Tabelle 5.17: LID-Leistungsverbesserungen durch Integration von höherem Wissen

5.8 Nachbehandlung durch das Sprachmodell

In der Literatur wurde für phonologische Ansätze vorgeschlagen, das linguistische Wissen als Nachbearbeitungsschritt auf die beste akustische Hypothese anzuwenden statt es in den Dekodierungsprozeß zu integrieren. Es wurden dazu von Zissmann [40] Untersuchungen veröffentlicht, die aber keinen signifikanten Unterschied bei den phonologischen Ansätzen ergaben. Der Gedanke der Nachbehandlung der besten Hypothese kann auf höheres linguistisches Wissen übertragen werden. Statt also das sprachliche Modell in die Dekodierung mit aufzunehmen, wird zunächst mit dem System CDohneLM die beste Hypothese berechnet und in einem anschließenden Nachbehandlungsschritt für diese beste Hypothese die Bewertung durch das sprachliche Modell ermittelt. Der Grundannahme ist, daß der Erkenner der tatsächlich gesprochenen Sprache auf der Testäußerung grammatikalisch wahrscheinlichere Hypothesen liefern wird als der Erkenner der fremden Sprache. Dieses System wird im folgenden als CDpostLM bezeichnet.

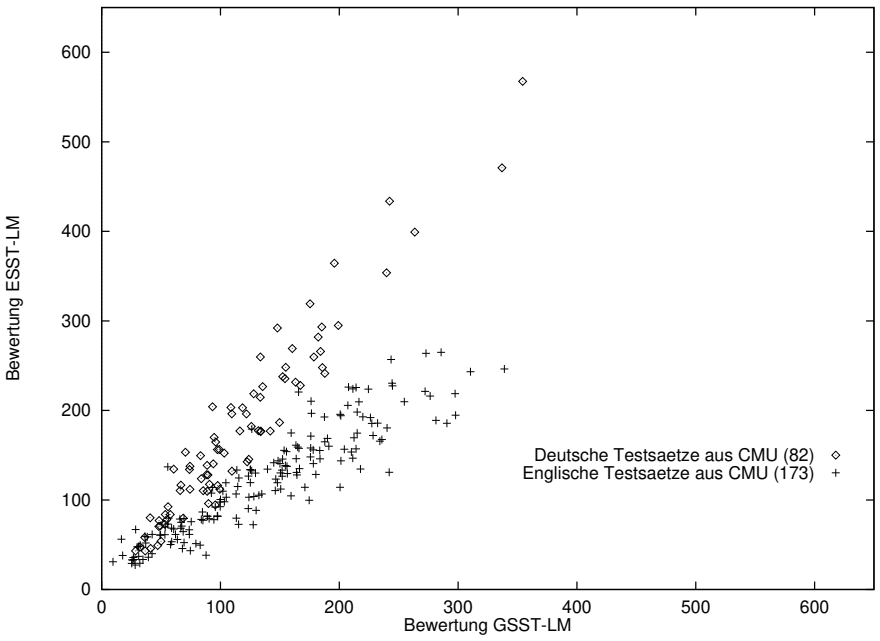

Abbildung 5.14: Nachbehandlung mit Sprachmodell auf CMU Daten

Auch für diese Experimente wurde zur Berechnung der Trenngeraden ein Perzeptron verwendet. Dies scheint auf den ersten Blick wenig plausibel, da die Berechnung von grammatikalischen Scores bei der Verwendung von Bigrammen auf Wahrscheinlichkeiten basiert. Es besteht somit kein Grund, nicht die Winkelhalbierende als Trenngerade zu nehmen. Die Verschiedenartigkeit der Grammatiken für die Sprachen Deutsch und Englisch, die sich in unterschiedlichen Perplexitäten und ungleich großem Vokabular ausdrückt, machen eine Normierung dennoch notwendig. Die Abbildung 5.14 zeigt die Resultate dieser Nachbehandlung der besten akustischen Hypothese durch das Sprachmodell für die Testäußerungen, die an der CMU aufgenommen wurden. Man sieht deutlich, daß das Sy-

stem CDpostLM die Testäußerungen mit wachsender Satzlänge sehr gut trennt. Für sehr kurze Sätze ist die Trennung nicht so gut. Das ist plausibel, denn längere Sätze enthalten mehr Worte und mit wachsender Zahl von Worten greift das Sprachmodel besser. Insofern sind die in Tabelle 5.18 dargestellten Ergebnisse irreführend, weil sie die Leistung über alle Satzlängen zusammengefaßt beschreiben. Durch weitere Experimente wurde eine interessante Beobachtung gemacht: durch die Eliminierung aller Geräusche aus der besten akustischen Hypothese vor der Nachbehandlung konnte die Identifizierungsleistung signifikant verbessert werden. Die Abbildung 5.15 zeigt diese Verbesserung wegen der Übersichtlichkeit wiederum nur für die CMU-Daten.

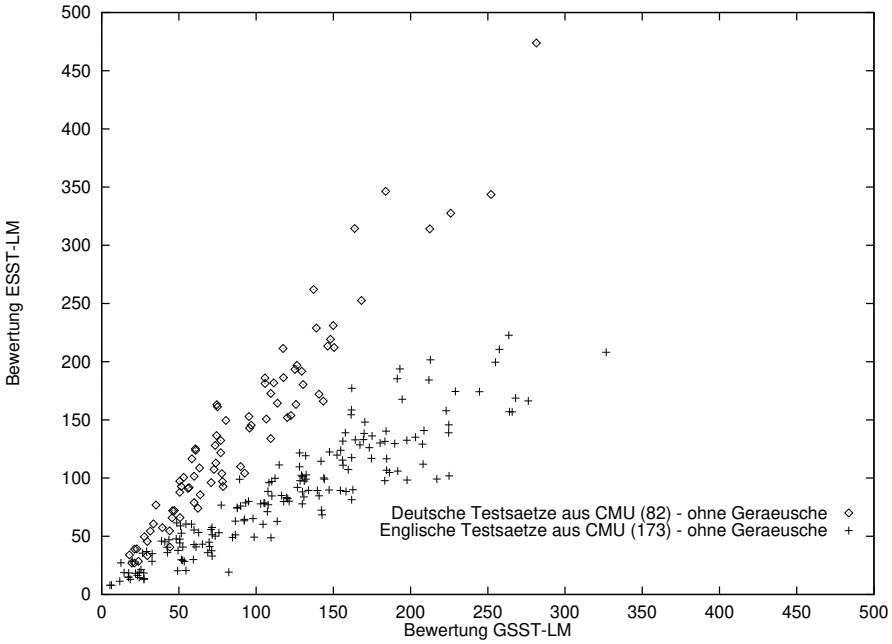

Abbildung 5.15: CDpostLM nach Geräuscheliminierung auf CMU Daten

System	Deutsch	Englisch	Gesamt
Aufnahmen in Karlsruhe			
CDpostLM	83,8%	81,3%	83,0%
CDpostLM ohne Geräusche	97,9%	74,7%	89,9%
Aufnahmen an der CMU			
CDpostLM	98,8%	72,3%	80,8%
CDpostLM ohne Geräusche	98,8%	91,9%	94,1%

Tabelle 5.18: LID-Leistung durch Nachbearbeitung mit Sprachmodell

Wie in der Abbildung deutlich wurde, wird die Trennbarkeit der Sprachen mit wachsender Satzlänge immer besser. Die Länge eines Satzes korreliert mit der Anzahl der gespro-

chenen Worte. Da Wortbigramme nur dann sinnvoll einsetzbar sind, wenn eine Äußerung mindestens zwei Worte enthält, sollten nicht alle Sätze in die Analyse mit aufgenommen werden, oder die LID-Leistung in Abhängigkeit der Anzahl Worte dargestellt werden.

System	Deutsch	Englisch	Gesamt
Aufnahmen in Karlsruhe			
alle Hypothesen	97,9% 3(142)	74,7% 19(75)	89,9%
mehr als 0 Worte	97,9% 3(142)	75,7% 18(74)	90,3%
mehr als 3 Worte	98,6% 2(141)	75,7% 18(74)	90,7%
mehr als 4 Worte	98,6% 2(141)	76,7% 17(73)	91,1%
mehr als 7 Worte	98,6% 2(141)	77,8% 16(72)	91,6%
Aufnahmen an der CMU			
alle Hypothesen	98,8% 1(82)	91,9% 14(173)	94,1%
mehr als 1 Wort	98,8% 1(82)	93,0% 12(171)	94,9%
mehr als 2 Worte	98,8% 1(82)	94,1% 10(169)	95,6%
mehr als 6 Worte	98,8% 1(82)	95,2% 8(167)	96,4%

Tabelle 5.19: Wortabhängige LID-Leistung druch Nachbearbeitung mit Sprachmodell

Die Tabelle 5.19 zeigt die LID-Leistungen des Systems CDpostLM in Abhängigkeit der Wortzahl. Bei der Betrachtung sollte man allerdings bedenken, daß die prozentualen Verbesserungen wegen des begrenzten Testmaterials zuweilen durch einzelne Äußerungen zustande kommen. Die Anzahl der Worte bezieht sich bei den Karlsruher Daten auf tatsächlich gesproche Worte, wie sie aus den Transkriptionen zu entnehmen waren. Bei den CMU-Daten standen keine Transkriptionen zur Verfügung, so daß es sich bei diesen Wortzahlangaben um die hypothetisierten Worte handelt. Die tatsächliche Zahl der Worte wird bei den deutschen Sätzen im Mittel um 4,3 Worte und relativ zur Gesamtzahl von Worten pro Äußerung betrachtet um 19% überschätzt. Bei den englischen Sätzen werden ebenfalls im Schnitt 4,3 Worte zuviel hypothetisiert, wegen der geringeren Anzahl Worte pro Äußerung liegt der relative Anteil mit 22,5% etwas höher. Legt man insgesamt ein Mindestvorkommen von 3 Worten zugrunde, dann wird durch das System CDpostLM die Leistung der sprachidentifizierenden Einheit bei den Karlsruher Daten um 16%, bei den CMU Daten um 34% gesteigert. Durch den Ausschluß der Hypothesen mit weniger als 3 Worten konnte für das System CDmitLM keine Leistungsverbesserung erzielt werden. Es handelt sich somit beim System CDpostLM um eine echte Steigerung der LID-Leistung.

5.9 Grammatikalisch erzwungene Akustik

In gewisser Weise wird der Verlauf des Viterbipfades aus der Sicht der reinen Akustik durch die Anwendung eines Sprachmodelles durch Abschnitte gezwungen, die akustisch gesehen nicht so attraktiv sind. Der Pfad geht dennoch diesen Weg, weil die Grammatik

System	Deutsch	Englisch	Gesamt
Aufnahmen in Karlsruhe			
CDzwang	88,7%	84,0%	87,1%
Aufnahmen an der CMU			
CDzwang	90,2%	83,8%	85,9%

Tabelle 5.20: LID-Leistung mit CDzwang

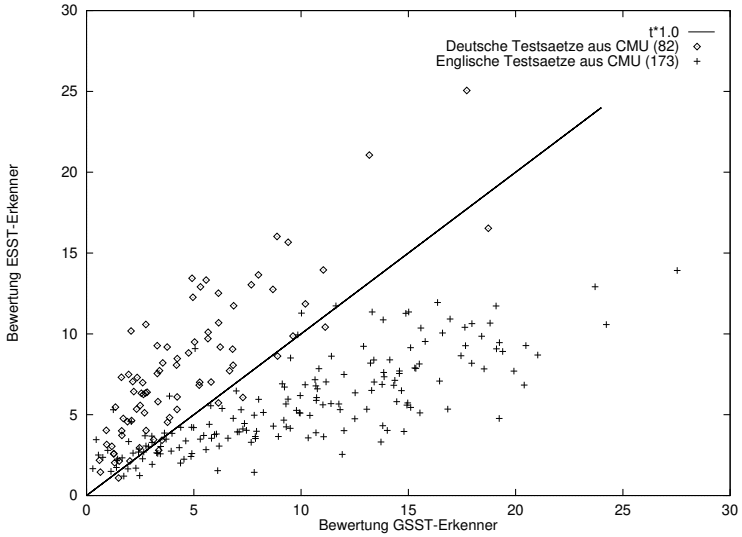

Abbildung 5.16: LID-Leistung mit CDzwang am Aufnahmeort CMU

ihn für plausibler hält. Diese Tatsache kann für die Identifizierung einer Sprache genutzt werden, wenn man davon ausgeht, daß bei einem System, welches seine eigene Sprache erkennt, die Akustik und die Grammatik besser harmonieren. Harmonie bedeutet in diesem Sinn, daß die Grammatik das Akustikmodell seltener durch schlecht passende Abschnitte zwingt als es das System der fremden Sprache tun würde. Zu diesem Zweck berechnet man die Differenz zwischen der akustischen Bewertung des Laufs ohne Sprachmodell CDohneLM und der akustischen Bewertung des Laufs mit Sprachmodell CDmitLM. Die Scoredifferenz CDmitLM - CDohneLM sollte für denjenigen Erkenner kleiner sein, der auf der Sprache der vorgelegten Äußerung trainiert wurde. Die Vorgehensweise wird mit CDzwang bezeichnet. Die Ergebnisse in Tabelle 5.20 bestätigen die Richtigkeit der Annahme. Immerhin kann man mit diesem Verfahren die gesprochene Sprache mit 85,9% bzw. 87,1% korrekt identifizieren. Die Abbildung 5.16 stellt die Ergebnisse exemplarisch für die Aufnahmen an der CMU dar. Die Entscheidungsgrenze ist mit eingezeichnet. Insgesamt konnte durch das Verfahren CDzwang jedoch nicht die Leistung der bisher vorgestellten Systeme übertroffen werden. Die Vorgehensweise wäre in der Praxis auch nur mit einem Mehraufwand an Rechenzeit durchführbar, da zwei beste Pfade jeweils einer mit und einer ohne Grammatik berechnet werden müssen.

Die Grafik 5.17 zeigt abschließend die Leistung der neu entwickelten Systeme CDohneLM, CDmitLM und das erfolgreichste System CDpostLM (dargestellt für Hypothesen mit mehr als 3 Worten) im Vergleich mit den aus der Literatur entnommenen bisher bekannten Systemen PohneLM und PmitLM, die kein höheres Wissen in den Identifizierungsprozeß integriert haben. Die Grafik enthält die Resultate getrennt nach Aufnahmeorten. Um nocheinmal den Einfluß des Kanals auf die Identifizierungsleistung zu verdeutlichen, werden die Resultate auch für die kanalabhängigen Daten gezeigt.

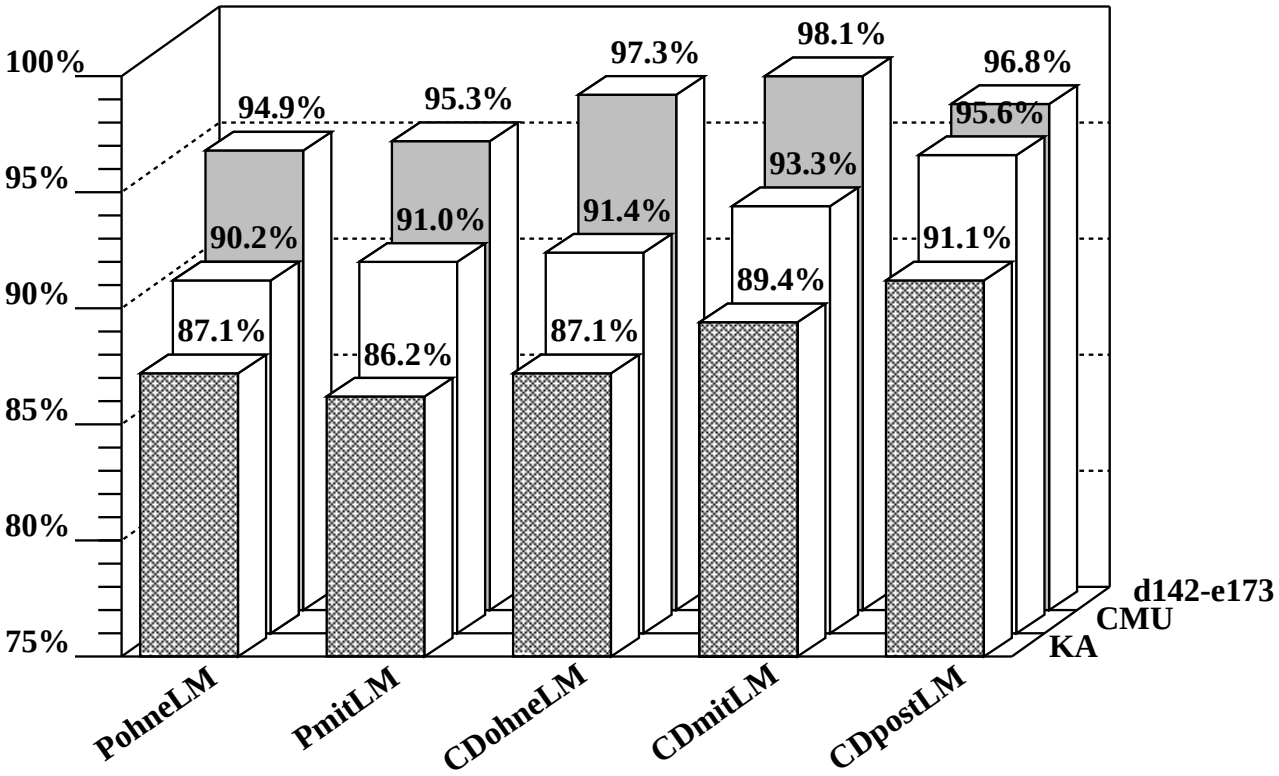

Abbildung 5.17: Leistungsvergleich aller vorgestellten Systeme

5.10 Die Unterscheidung der Sprachen Deutsch, Englisch und Spanisch

Spanische Sprachdaten standen nur vom Aufnahmeort CMU zur Verfügung. Um die Wirkung der einzelnen Verfahren aber auf beiden Aufnahmeorten vorführen zu können, wurden die Experimente bisher nur für den Zweisprachenfall Englisch-Deutsch vorgestellt. Daher erscheinen die Experimente auf allen drei Sprachen nun in diesem letzten Abschnitt gesondert.

System	CMUd142 - e173	CMUd142 - s86	s86 - e173
alle Sprachdaten aufgenommen an der CMU			
PohneLM	90,2%	70,2%	91,9%
PmitLM	91,0%	74,9%	89,9%
CDohneLM	91,4%	82,1%	96,5%
CDmitLM	93,3%	88,6%	97,7%
CDpostLM ohne Geräusche	94,1%	95,2%	90,3%

Tabelle 5.21: LID-Leistung aller Systeme für Deutsch, Englisch und Spanisch

Die Tabelle 5.21 zeigt für alle drei Sprachen die Identifizierungsleistungen für alle vorgestellten Systeme. Es wurden nur die Daten an der CMU verglichen. Es ergeben sich drei Zweisprachentests, je ein Test Deutsch-Spanisch, Spanisch-Englisch und den bereits ausführlich dargestellten Test Deutsch-Englisch, der in der Tabelle zum Vergleich nocheinmal aufgeführt ist. Die prozentualen Angaben sind bereits über alle getesteten Äußerungen gemittelt. Wie aus der Tabelle ersichtlich, ist die Unterscheidung Englisch-Spanisch einfacher als Deutsch-Englisch. Dies wurde bereits in anderen Untersuchungen berichtet. Gut zu erkennen ist auch die durch Lexikon und Grammatik erreichte Leistungssteigerung. Auffallend ist das schlechte Abschneiden des Systems CDpostLM für den Zweisprachentest Spanisch-Englisch. Am schwierigsten ist die Unterscheidung zwischen Deutsch und Spanisch. Man sieht aber auch hier sehr gut, daß die mäßige Leistung auf der akustischen und phonologischen Ebene durch den Einsatz höheren Wissens signifikant verbessert werden kann. Für die Identifizierung dieser zwei Sprachen bringt auch das System CDpostLM einen großen Leistungszuwachs. Für die Identifizierung von Spanisch kann man durch das System CDpostLM unter Ausschluß der Hypothesen, die weniger als 3 Worte enthalten keine Verbesserung erwarten, da unter den spanischen Testsätzen keine einzige Äußerung mit 3 Worten oder weniger ist. Insgesamt werden im Spanischen auch wesentlich mehr Worte hypothetisiert als beim Deutschen und Englischen. Im Mittel werden 6,9 Worte mehr hypothetisiert, das entspricht einem Anteil von 30% bezogen auf die tatsächlich gesprochenen Worte[3]. Im Spanischen werden somit wesentlich mehr Worte hypothetisiert

[3]Diese Angaben beziehen sich alle auf den Lauf mit CDohneLM, denn die von diesem System berechneten Hypothesen sind die Berechnungsgrundlage für das System CDpostLM

als in den Sprachen Deutsch und Englisch, bei denen der Schnitt etwa 20% beträgt, wie bereits in Abschnitt 5.8 dargestellt.
Die Abbildung 5.18 zeigt links die Unterscheidung der Sprachen Deutsch und Spanisch für das dafür am besten geeignete System CDpostLM mit einer Identifizierungsleistung von 95,2%. Die Abbildung zeigt rechts die Verhältnisse beim System CDmitLM für die Unterscheidung von Englisch und Spanisch.

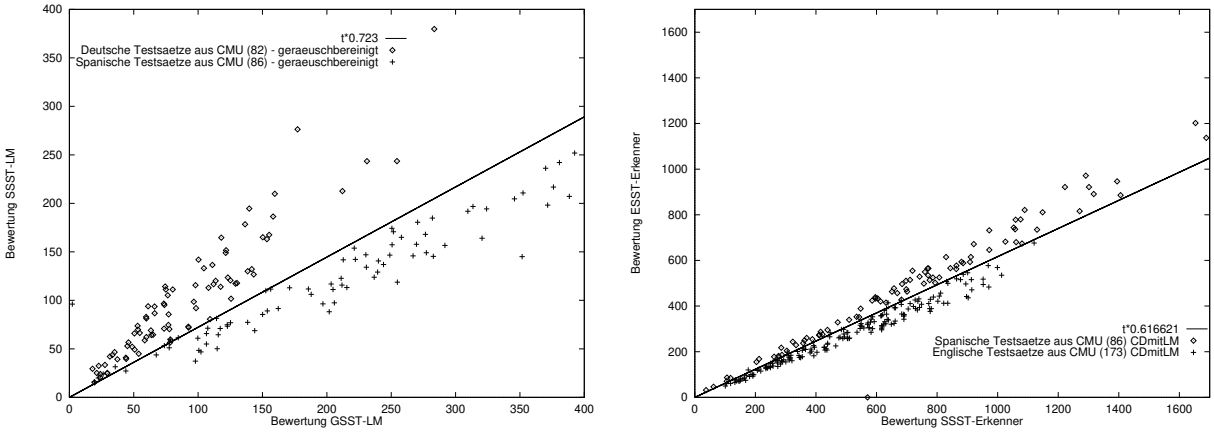

Abbildung 5.18: Unterscheidung Deutsch, Englisch und Spanisch

Abschließend wurde ein Dreisprachentest durchgeführt, d.h. das System muß eine aus drei möglichen Sprachen statt wie bisher eine aus zwei Sprachen korrekt identifizieren. Zur Normierung wurde hier die Kalibrierung der Bewertungen durch den erkennerspezifischen Scoreperframe vorgenommen um anschließend die *Maximum Likelihood*-Regel anwenden zu können. Der Dreisprachentest ergab eine Identifizierungsleistung von 84%. Diese Leistung ist signifikant schlechter als die erreichten Leistungen bei den Zweisprachentests. Dies liegt einerseits daran, daß die Trennung von 3 Sprachen ein schwierigeres Problem darstellt, als die Unterscheidung von 2 Sprachen. Andererseits wurde für den Dreisprachentest statt des Perzeptrons die Normierung durch Kalibrierung verwendet, was zu Leistungseinbußen führt, wie in Abschnitt 5.3 gezeigt worden war.

Kapitel 6

Diskussion und Ausblick

Das Ziel der vorliegenden Diplomarbeit war die Entwicklung eines sprachidentifizierenden Moduls für das Übersetzungssystem JANUS. Zur Lösung dieser Aufgabe wurden zunächst die bereits aus der Literatur bekannten Ansätze studiert und dargestellt. Zum Vergleich wurden zwei Ansätze jeweils ein Phonemerkenner ohne Sprachmodell und ein Phonemerkenner mit Phonembigrammen nachgebildet und deren Ergebnisse auf der Datenbasis SST bestimmt. Darüberhinaus wurden neue Wege zur Lösung des Identifizierungsproblems beschritten. Es wurde höheres Wissen in Form eines Lexikons und einer Grammatik auf Wortebene in das sprachidentifizierende Modul integriert. Diese Vorgehensweise setzt jedoch das Vorhandensein solcher Quellen und eine spracherkennende Einheit voraus. Der Aufwand der Erstellung solcher Wissensquellen eigens für die Identifizierung wäre zu aufwendig, und müßte zu recht als ineffizient abgelehnt werden. Es war jedoch eine Vorgabe dieser Arbeit, ein Modul zu entwickeln, daß als „front-end“ für JANUS eingesetzt werden soll. Da der Spracherkennungsprozeß ein Lexikon und eine Grammatik voraussetzt, waren damit bereits alle Wissensquellen verfügbar. Insofern muß die Einsatzmöglichkeit der hier vorgestellten Ansätze auf die Notwendigkeit der Kopplung mit einem Spracherkennungssystem eingeschränkt werden.

Im Prinzip ist das in dieser Arbeit vorgestellte beste System zur Identifizierung von Sprachen einfach aber sehr leistungsfähig. Als einzige neue Komponente muß eine Instanz in das JANUS-System integriert werden, die die Trenngerade bzw. die Normierungsfaktoren zur Scorenormierung berechnet. In dieser Arbeit wurden zwei Instanzen, die im Zuge der Test- oder Trainingsläufe mitgelernt werden können vorgestellt. Eine Methode besteht im Trainieren eines Neuronalen Netzes, wobei für den Zweisprachenfall ein einfaches Perzeptron zum Erlernen einer Trenngerade ausreicht. Für den Mehrsprachenfall genügt ein Perzeptron jedoch nicht mehr. Eine weitere Methode ist die Normierung der Scores durch eine Kalibrierung. Dabei würde das regelmäßige Abspeichern des aktuellen erkennerspezifischen Scoresperframe ausreichen um mit der einfachen *Maximum Likelihood*-Regel zwischen mehreren Sprachen unterscheiden zu können. Dieses einfaches Verfahren wird jedoch mit kleinen Einbußen in der Identifizierungsleistung erkauft.

Als erschreckend drastisch hat sich der Einfluß unterschiedlicher Kanaleigenschaften der Testdaten auf die Identifizierungsleistung herausgestellt. Um so mehr verwundert es, daß diesem Problem in der Literatur so wenig Aufmerksamkeit geschenkt wird. Die Experimente in dieser Arbeit ergaben, daß die Eliminierung der Kanalunterschiede den Identifizierungsfehler verdoppelt. Da man bei der Identifizierung von Sprachen vom Kanal unabhängig sein möchte, sollten Verfahren zur Kanaladaption angewendet werden.

Eine interessante Folgearbeit wäre die Analyse des Systems, für den Fall, daß noch weitere zu identifizierende Sprachen hinzugefügt werden. Mit wachsender Zahl zu unterscheidender Sprachen müssen immer mehr Erkenner parallel gestartet werden. Der Berechnungsaufwand wird damit größer und die Redundanz der Berechnungen steigt. Für ein kompaktes, multilinguales Übersetzungssystem ist dieser Effekt nicht erwünscht. Es muß daher nach Strategien gesucht werden, die die Redundanz der Rechenarbeit verringern. Eine Möglichkeit wäre ein integrales System mit gemeinsamer Nutzung ähnlicher Phoneme. Eine weitere Möglichkeit wäre die Konzeption eines mehrstufigen Verfahrens, bei dem zunächst grob Sprachgruppen unterschieden werden und anschließend aus der gewählten Sprachgruppe mit feineren Verfahren die korrekte Sprache identifiziert wird. Zu untersuchen wäre dann natürlich, ob geeignete Sprachgruppen existieren.

Ein weiteres Problem ergibt sich bei fehlendem apriori Wissen über die zu unterscheidenen Sprachen. Die Vergleiche der Resultate für die Zweisprachentests Englisch-Spanisch, Deutsch-Spanisch und Deutsch-Englisch haben gezeigt, daß ein System, welches ein Sprachenpaar erfolgreich unterscheiden konnte, nicht unbedingt für ein anderes Sprachenpaar genauso gut funktioniert. Wenn die zu unterscheidenen Sprachen apriori nicht bekannt sind, kann das Wissen, welches der Systeme für die Unterscheidung der vorliegenden Sprachen am besten funktioniert, nicht eingesetzt werden. Dasselbe Problem ergibt sich im Mehrsprachenfall, falls für die beteiligten Sprachen unterschiedliche Systemverhalten festgestellt wurden, wie es in dieser Arbeit für die Sprachen Deutsch, Englisch und Spanisch festgestellt werden konnte. In einem solchen Fall müssen andere Strategien angewendet werden. Denkbar wäre beispielsweise die parallele Verfolgung aller Ansätze und anschließende Entscheidung für den besten Ansatz. Dies könnte wiederum durch Neuronale Netze oder Wahrscheinlichkeiten entschieden werden.

Abschließend kann festgehalten werden, daß die im Rahmen dieser Diplomarbeit entwickelten Ansätze die gestellte Aufgabe sehr gut lösen, daß aber für den Einsatz des sprachidentifizierenden Moduls in einem größeren Kontext noch weitere Forschungsarbeit geleistet werden muß.

Literaturverzeichnis

[1] M. Abe, K. Shikano, and H. Kuwabara: *Statistical Analysis of Bilingual Speaker's Speech for Cross-language Voice Conversation* in: Journal of the acoustical Society of Amerika (JASA) 90(1), S. 76-82, 1991.

[2] O. Andersen, P. Dalsgaard, and W. Barry: *Data-Driven Identification of Poly- and Mono-Phonemes for four European Languages* in: Proc. Eurospeech, S. 759-762, Berlin 1993.

[3] O. Andersen, P. Dalsgaard, and W. Barry: *On the Use of Data-Driven Clustering Techniques for Language Identification of Poly- and Mono-phonemes for four European Languages* in: Proc. ICASSP, S. 121-124, volume1, Adelaide 1994.

[4] K. Atkinson: *Language Identifikation from nonsegmental cues* in: Journal of the Acoustic Society of America, 44:378(A), 1968.

[5] L.R. Bahl, F. Jelinek, R.L. Mercer: *A Maximum Likelihood Approach to Continuous Speech Recognition* in: A. Waibel and K.-F. Lee (editors) Readings in Speech Recognition, S. 308-319, Morgan Kaufmann 1990.

[6] W. Barry and P. Dalsgaard: *Speech Database annotation. The Importance of a Multi-Lingual approach* in: Proc. Eurospeech, S. 13-20, volume 1, Berlin 1993.

[7] K.M. Berkling, T. Arai, and E. Barnard: *Analysis of Phoneme-based Features for Language Identification* in: Proc. ICASSP, S. 289-292, volume 1, Adelaide 1994.

[8] K.M. Berkling and E. Barnard: *Language Identification of Six Languages Based of a Common Set of Broad Phonemes* in: Proc. ICSLP, S. 1891-1894, Yokohama 1994.

[9] P. Dalsgaard and O. Andersen: *Application of Inter-Language Phoneme Similarities for Language Identification* in: Proc. ICSLP, S. 1903-1906, Yokohama 1994.

[10] T.J. Hazen and V.W. Zue: *Automatic Language Identification using a Segment-based Approach* in: Proc. Eurospeech, S. 1303-1306, Berlin 1993.

[11] T.J. Hazen and V.W. Zue: *Recent Improvements in an Approach to Segment-based Automatic Language Identification* in: Proc. ICSLP, S. 1883-1886, Yokohama 1994.

[12] X.D. Huang and M.A. Jack: *Semi-continuous Hidden Markov Models for Speech Signals* in: A. Waibel and K.-F. Lee (editors) Readings in Speech Recognition, S. 340-346, Morgan Kaufmann 1990.

[13] S. Itahashi, J.X. Zhou, and K. Tanaka: *Spoken Language Diskrimination using Speech Fundamental Frequency* in: Proc. ICSLP, S. 1899-1902, Yokohama 1994.

[14] F. Jelinek: *Statistical Techniques of Speech and Lanuguage Modeling* in: bisher unveröffentlichtes Manuskript vom Autor persönlich überreicht, Nov. 1994.

[15] S. Kadambe and J.L. Hieronymus: *Spontaneous Speech Language Identification with a Knowledge of Linguistics* in: Proc. ICSLP, S. 1879-1882, Yokohama 1994.

[16] T. Kemp: *Data-Driven Codebook Adaption in phonetically tied SCHMMS* to appear in Proc. ICASSP, Detroit 1995.

[17] L.F. Lamel and J. Gauvain: *Identifying Non-linguistic Speech Features* in: Proc. Eurospeech, S. 23-30, Berlin 1993.

[18] L.F. Lamel and J. Gauvain: *Language Identification using Phone-based Acoustic Likelihoods* in: Proc. ICASSP, S. 305-308, volume 1, Adelaide 1994.

[19] K.P. Li: *Automatic Language Identification using Syllabic Features* in: Proc. ICASSP, S. 297-300, volume 1, Adelaide 1994.

[20] Y.K. Muthusamy, E. Barnard, and R.A. Cole: *Reviewing Automatic Language Identification* in: IEEE Signal Processing Magazin, Vol. 11 Nr. 4, S. 33-41, Oktober 1994.

[21] Y. Muthusamy, K. Berkling, T. Arai, R.A. Cole, and E. Barnard: *Comparison of Approaches to Automatic Language Identification using Telephone Speech* in: Proc. Eurospeech, S. 1307-1310, Berlin 1993.

[22] Y.K. Muthusamy, N. Jain, and R.A. Cole: *Perceptual Benchmarks for Automatic Language Identification* in: Proc. ICASSP, S. 333-336, Adelaide 1994.

[23] Y.K. Muthusamy and R.A. Cole: *Automatic Segmentation and Identification of ten Languages using Telephone Speech* in: Proc. ICSLP, S. 1007-1010, Banff 1992.

[24] H.H. Nagel: *Material zur Vorlesung Kognitive Systeme* Institut für Algorithmen und Kognitive Systeme, Fakultät für Informatik der Universität Karlsruhe (TH), 1991.

[25] H. Ney: *Modeling and Search in Continuous Speech Recognition* in: Proc. Eurospeech, Vol. 1 S. 491-498, Berlin 1993.

[26] L.R. Rabiner: *A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition* in: A. Waibel and K.-F. Lee (editors) Readings in Speech Recognition, S. 267-296, Morgan Kaufmann 1990.

[27] P. Ramesh and D.B. Roe: *Language Identification with Embedded Word Models* in: Proc. ICSLP, S. 1887-1890, Yokohama 1994.

[28] D.R. Reddy: *Speech Recognition by Machine: A Review* in: A. Waibel and K.-F. Lee (editors) Readings in Speech Recognition, S. 8-30, Morgan Kaufmann 1990.

[29] A.A. Reyes, T. Seino, and S. Nakagawa: *Three Language Identification methods based on HMMs* in: Proc. ICSLP, S. 1895-1898, Yokohama 1994.

[30] M. Savic, E. Acosta, and S.K. Gupta: *An Automatic Language Identification System* in: Proc. ICASSP, S.817-820, Toronto 1991.

[31] P. Scheytt: *Wordspotting-Techniken* Studienarbeit, unveröffentlicht, Universität Karlsruhe und Politecnico Di Torino 1995.

[32] T. Schultz, and I. Rogina: *Acoustic and Language Modeling of Human and Nonhuman Noises for Human-to-Human Spontaneous Speech Recognition* to appear in: Proc. ICASSP, Detroit 1995.

[33] T. Seino and S. Nakagawa: *Spoken Language Identification using Ergodic HMM with Emphasized State Transition* in: Proc. Eurospeech, S. 133-136, volume 1, Berlin 1993.

[34] M. Sugiyama: *Automatic Language Recognition using Acoustic Features* in: Proc. ICASSP, S. 813-816, Toronto 1991.

[35] B. Suhm, P. Geutner, T. Kemp, A. Lavie, L. Mayfield, A.E. McNair, I. Rogina, T. Sloboda, W. Ward, M. Woszczyna, and A. Waibel: *JANUS: Towards Multilingual Spoken Language Translation* in: DARPA Speech and Natural Language Workshop, 1995.

[36] R.C.F Tucker, M.J. Carey, and E.S. Parris: *Automatic Language Identification using subword models* in: Proc. ICASSP, S.301-304, volume 1, Adelaide 1994.

[37] M. Woszczyna, N. Aoki-Waibel, F.D. Buø, N. Coccaro, K. Horigushi, T. Kemp, A. Lavie, A. McNair, T. Polzin, I. Rogina, C.P. Rose, T. Schultz, B. Suhm, M. Tomita, and A. Waibel: *Janus 93: Towards Spontaneous Speech Translation* in: Proc. ICASSP, S. 345-349, Adelaide 1994.

[38] M. Woszczyna, N. Coccaro, A. Eisele, A. Lavie, A. McNair, T. Polzin, I. Rogina, C.P. Rose, T. Sloboda, M. Tomita, J. Tsutsumi, N. Aoki-Waibel, A. Waibel, W.Ward: *Recent Advances in JANUS: A Speech Translation System* in: Proc. Eurospeech, S. 1295-1298, Berlin 1993.

[39] M.A. Zissmann: *Automatic Language Identification using Gaussian Mixtures and Hidden Markov Models* in: Proc. ICASSP, S. 309-402, volume 2, Minneapolis 1993.

[40] M.A. Zissmann and E. Singer: *Automatic Language Identification of Telephone Speech Messages using Phoneme Recognition and N-gram Modeling* in: Proc. ICASSP, S. 305-308, volume 1, Adelaide 1994.

Hiermit erkläre ich an Eides Statt, daß ich die vorliegende Diplomarbeit selbständig und ohne unzulässige fremde Hilfe angefertigt habe. Die verwendeten Literaturquellen sind im Literaturverzeichnis vollständig aufgeführt.

Heidelberg, den 30. 4. 1995 .....

Tanja Schultz
Dossenheimer Landstraße 86/1
69121 Heidelberg