Automatische Identifizierung spontan gesprochener Sprachen mit neuronalen Netzen

Tanja Schultz und Hagen Soltau
Interactive Systems Labs, ILKD
Universität Karlsruhe, 76128 Karlsruhe

Abstract

Automatic language idenfication (LID) is one of the keyproblems in building multilingual speech recognition and translation systems. In this paper we present a front-end module to identify one out of four languages German, English, Spanisch and Japanese for use in spontaneous speech-to-speech translation systems like JANUS. We compare two different approaches to classify the speech data, on the one hand we used the maximum likelihood method on the other hand neuronal networks. It can be shown, that neuronal nets leads to better results which is overall 86% for the four language test.

Die automatische Identifikation von Sprachen (LID) ist ein Problem, dessen Lösung für die Entwicklung multilingualer Spracherkennungs- und Übersetzungssysteme zwingend erforderlich ist. In dieser Arbeit stellen wir ein Modul vor, das automatisch eine aus den vier Sprachen Deutsch, Englisch, Spanisch und Japanisch identifiziert. Dieses Modul wurde für das multilinguale Sprache-zu-Sprache Übersetzungssystem JANUS entwickelt. Zur Klassifikation des vorliegenden spontan gesprochenen Datenmaterials wurden zwei Ansätze miteinander verglichen, die Maximum Likelihood Methode und neuronale Netze. Die Ergebnisse zeigen, daß neuronale Netze der ML-Methode in dieser Anwendung überlegen sind. Beim 4-Sprachentest wird eine Identifikationsleistung von 86% erreicht.

1 Einleitung

Die zunehmende internationale Verflechtung von Wirtschaft, Politik und Gesellschaft erhöht den Bedarf an Einrichtungen und Anwendungen, die eine schnelle und problemlose Kommunikation im globalen Ausmaß ermöglichen. Daher sind Systeme, die Dienstleistungen auf der Basis natürlich gesprochener Sprache anbieten und dabei Sprachbarrieren überwinden, von besonderem Interesse.

Solche Kommunikationssysteme sollten nicht nur Sprache verarbeiten, erkennen und verstehen können, sondern auch die Fähigkeit der Multilingualität besitzen. Als *multilingual* werden hier automatische Sprachsysteme bezeichnet, die als Ein- und/oder Ausgabesprache verschiedene Sprachen tolerieren. Denkbare Anwendungen solcher Sprachsysteme sind multilinguale Übersetzungssysteme, Informationsdienste in öffentlichen Institutionen wie etwa Zugauskunftsdienste, Messeinformations- oder Flughafenserviceeinrichtungen sowie telefonbasierte Dienste wie Vermittlung von Ferngesprächen, Notrufeinrichtungen, Rufnummernauskünfte und Hotelreservierungen. Sofern diese multilingualen Sprachapplikationen fremdsprachlichen Benutzern zugänglich gemacht werden sollen, müssen sie die verwendete Eingabesprache automatisch identifizieren können.
Unser Ziel ist es, ein sprachenidentifizierendes Modul für das System JANUS zu entwickeln. JANUS ist ein multilinguales Sprache-zu-Sprache Übersetzungssystem, das spontan gesprochene Äußerungen in den Sprachen Deutsch, Englisch, Spanisch oder Japanisch erkennt, diese anschließend in eine dieser vier Sprachen übersetzt und via Sprachsynthese ausgibt. Die Aufgabe der automatischen sprachenidentifizierenden Einheit besteht darin, anhand der Vorgabe eines möglichst kurzen Sprachabschnittes die tatsächlich gesprochene Sprache mit möglichst hoher Präzision zu bestimmen.

2 Stand der Forschung

Forschung auf dem Gebiet der automatischen Identifizierung von Sprachen (LID) wird schon seit 20 Jahren betrieben. Die erste bekannte englischsprachige Arbeit stammt von Atkinson aus dem Jahr 1968. Die Untersuchungen aus den frühen Jahren sind allerdings nicht sehr zahlreich und setzten aus zwei Gründen wenig Impulse: ersten fehlten die Hinweise auf experimentelle Details und zweitens gab es bis zum Jahr 1992 keine allgemein zugängliche multilinguale Datenbasis, auf deren Grundlage verschiedene Ansätze evaluierbar gewesen wären. Erst seit in jüngster Zeit Spracherkennungssysteme auf der Basis großer Wortschätze für viele verschiedenen Fremdsprachen entwickelt werden, erwacht das Interesse an LID erneut. Die Erstellung multilingualer Datenbasen wie etwa der OGI-Korpus [4] und die Datenbasis SST ermöglicht es nun, verschiedene Forschungsansätze miteinander zu vergleichen.
Die Arbeiten zum Thema LID unterscheiden sich an der Anzahl der zu identifizierenden Sprachen (zwei bis elf), anhand des Sprachdatenmaterials, das herangezogen wird (isolierte Einheiten, gelesene Sprache, spontane Sprache), nach den verwendeten Klassifikationsmethoden (neuronale Netze, HMMs, Vektorquantisierung), und nach den zur Identifizierung verwendeten Sprachenmerkmalen (Lautmerkmale, Prosodie, segment- oder silbenbasierte Information).
In einer Studie von Muthusamy et. al. [5] in der eine aus 10 Sprachen identifiziert werden sollte, gaben die Versuchspersonen an, daß sie eine Kombination

von "'Phonem"'- und "'Wordspotting"', sowie die phonetischen und prosodischen Merkmale einer Sprache zur Identifizierung verwenden. Wir unterscheiden zwischen den folgenden fünf Informationsquellen, die zur Spracheniden-tifizierung genutzt werden könnten: das Lautinventar einer Sprache (im folgenden als akustisch-phonetische Merkmale bezeichnet), die Kombinationsmöglichkeiten von Lautfolgen (im folgenden phonologische Merkmale einer Sprache genannt), prosodische Merkmale wie Betonung, Intonation, Rhythmus, Tempo und Pausen, sowie das Aussprachelexikon (Wortschatz und Aussprache der Worte) und die grammatikalische Struktur einer Sprache.
Aufgrund der engen Verbindung zwischen LID und Spracherkennung werden in einem Großteil der Forschungsarbeiten komplette Spracherkennungssysteme zur Identifizierung von Sprachen verwendet. Die eigentlich erkannte Sequenz wird als Information zur Identifikation dabei nicht betrachtet, wichtig ist nur die Sprache in der die Sequenz vorliegt. In den meisten Forschungsarbeiten wird bisher ausschließlich akustisch-phonetisches Wissen auf der Grundlage von Einheiten wie Phonemen [9], [2] oder Phonemklassen [3] herangezogen. Einige Autoren fügen phonologisches Wissen in Form von Phonembigrammen [2] oder -trigrammen [1] hinzu. Die meisten Untersuchungen beschränken sich allerdings auf den Einsatz von phonembasierten Wissen, da der Rechenbedarf und der Aufwand für die Erstellung dieser Wissensquellen niedrig ist. Prosodische Wissensquellen werden bisher nur in sehr wenigen Arbeiten eingesetzt und erzielen keine signifikante Verbesserung der Identifizierungsleistung [1]. Wird die Identifikation von Sprachen als "'Vorverarbeitung"' zur Erkennung und Übersetzung von Sprache betrieben, wie es bei multilingualen Übersetzungssystemen der Fall ist, kann die Information der erkannten Sequenz weiterverwendet werden. Daher lohnt sich dort der Einsatz von höheren Wissensquellen wie Aussprachelexikon und Grammatik, der mit einem Mehraufwand für deren Erstellung und an Rechenbedarf für die Suchalgorithmen verbunden ist. Unsere frühere Experimente haben gezeigt, daß durch die Integration wortbasierter Lexika und Grammatiken in Form von statistischen Language Modellen eine wesentliche Verbesserung der Sprachenidentifizierungsleistung erreicht werden kann [6], [7].

3 Die multilinguale Datenbasis SST

Zur Entwicklung des hier beschriebenen Identifizierungssystems wird die multilinguale Datenbasis SST (**S**pontaneous **S**cheduling **T**ask) spontan gesprochener Dialoge verwendet. Dieser multilinguale Korpus enthält Terminabsprachen zwischen zwei Gesprächspartnern in den Sprachen Deutsch, Englisch, Spanisch und Japanisch. Große Teile des deutschsprachigen Materials, einige Teile des englischen Materials sowie das gesamte japanische Sprachdatenmaterial wurden im Rahmen des BMBF-Verbundprojektes VERBMOBIL gesammelt und transkribiert. Die deutschen Dialoge wurden an vier verschiedenen Aufnahmeorten (Kiel, Bonn, München und Karlsruhe) gesammelt, die englischen und spanischen

Dialoge wurden in Pittsburgh (USA) aufgenommen. Die japanischen Dialoge wurden am Advances Telephonic Research Laboratory in Kyoto (Japan) und an der University of Electro-Telecommunication in Tokyo (Japan) aufgezeichnet. Die Tabelle zeigt das zur Verfügung stehende Material für alle vier Sprachen.

Sprachen	Äusserungen	Stunden
Deutsch	12292	30.5
Englisch	7644	6.9
Spanisch	5730	10.7
Japanisch	3311	8.0

Table 1: Der multilinguale Korpus SST

4 Systemarchitektur

Grundsätzlich unterscheidet man bei sprachenidentifizierenden Systemen auf der Basis von Spracherkennern zwei Architekturformen. Bei der *integralen* Architektur wird ein einziger globaler Spracherkenner für alle zu identifizierenden Referenzsprachen trainiert. In diesem Modell sind alle sprachenspezifischen Eigenheiten integriert und konkurrieren beim Dekodieren der Testäußerung miteinander (vgl. [3]). Bei der *parallelen* Architektur wird für jede zu unterscheidende Referenzsprache ein eigenständiger Spracherkenner trainiert. Bei der Identifizierung laufen diese Spracherkenner parallel und erzeugen beim Dekodieren der unbekannten Testäußerung für jede Sprache eine Bewertung (Score). Diejenige Sprache, deren Erkenner den besten Score für die Testäußerung ermittelt hat, wird als die gesprochene Sprache identifiziert. Die parallele Architektur wird von einem Großteil der Forscher verwendet (vgl. [2], [1], [8]).
Die integrale Architektur hat den Nachteil, daß mit zunehmender Zahl der zu identifizierender Sprachen die Anzahl der zu integrierenden sprachenspezifischen Einheiten steigt. Dadurch wird erstens die Klassifizierung wegen wachsender Ambiguitäten schwieriger, zweitens steigt die Berechnungsdauer der Algorithmen an. Bei der parallelen Architektur kann man sehr verschiedene Erkenner einsetzen und deren Ausgaben parallel berechnen, mit der Zahl der Sprachen wächst aber der insgesamt zu leistende Rechen- und Speicherbedarf.

Die Abbildung 1 zeigt die Architektur eines sprachenidentifizierenden Systems. Im Fall einer integralen Architektur würden die Systemteile innerhalb des gestrichelt gezeichneten Rahmens zu einem einzigen gemeinsamen Erkenner verschmelzen. Die anschließende Entscheidung würde beim integralen Ansatz entfallen, da dies ein Bestandteil des integrierten Erkenners ist.

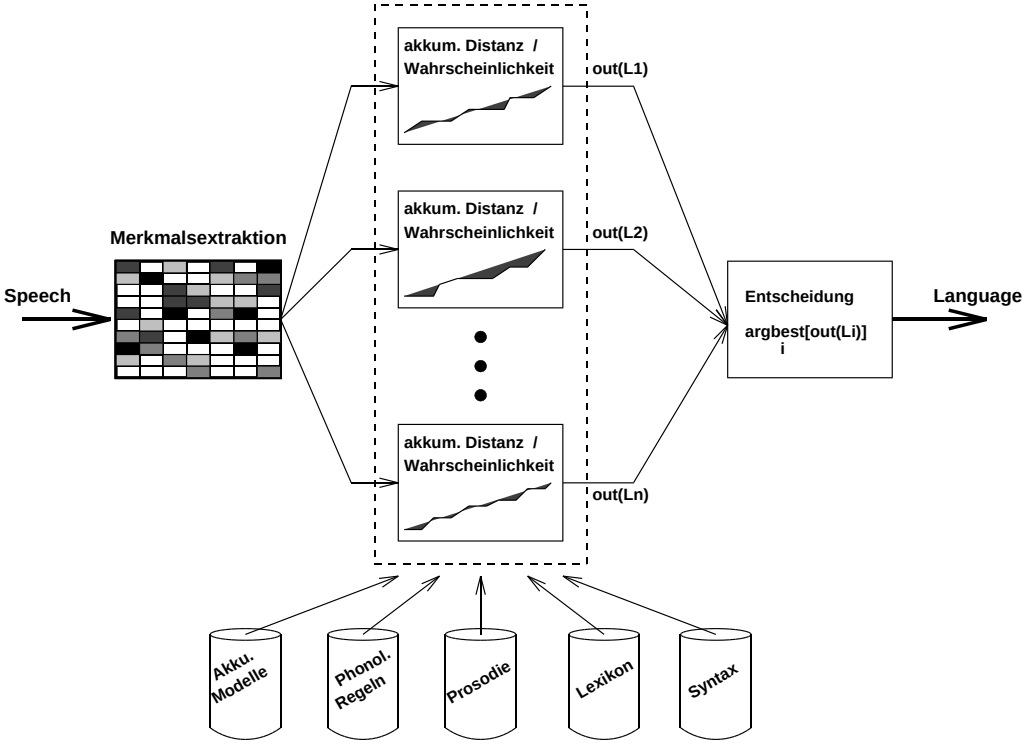

Figure 1: Architektur eines sprachenidentifizierenden Systems

5 Experimente

Für die Experimente wird die parallele Architektur verwendet. Wie die Abbildung 1 zeigt, erstellt man für jede der vier Sprachen einen eigenständiger Erkenner. Um die Bedeutung unterschiedlicher Wissensquellen zu erforschen, konzipierten wir zwei unterschiedliche Systeme P-3PT und W-3LM. Zum Trainieren der Erkenner und Berechnen der statistischen Language Modelle wird das Material aus Tabelle 1 abzüglich einer sprecherdisjunkten Testmenge verwendet.

System P-3PT
Das System P-3PT ist ein Phonemerkenner, der auf einer akustisch-phonetischen und phonologischen Wissensquelle aufbaut. Für jede der vier Sprachen wird ein eigenständiges P-3PT System trainiert, dessen kontextunabhängigen Phoneme durch SCHMMs mit 50 mixture Gaussians modelliert werden. Für die deutsche Sprache wird ein Erkenner mit 67 Phonemen, für das Englische mit 54 Phonemen, für Spanisch mit 48 Phonemen und für das Japanische mit 44 Phoneme verwendet. Die Phonemsätze enthalten jeweils spezielle

Geräuschmodelle zur Modellierung von nichtsprachlichen Ereignissen. Beim Dekodierungsprozess wird zusätzlich phonologisches Wissen in Form einer Phonem-Trigram-Grammatik angewendet. Dieses System ist sehr schnell und recheneffizient, allerdings ist die in der erkannten Phonemsequenz enthaltene Information für den nachfolgenden Übersetzungsprozeß nicht verwendbar. P-3PT eignet sich daher besonders für dedizierte schnelle Identifizierungssysteme.

System W-3LM
Das System W-3LM ist ein wortbasierter Erkenner, der ein Aussprachewörterbuch enthält. Dieses legt den Wortschatz fest und beschreibt auf welche Weise Phoneme zu Worten konkateniert werden können. Neben dem Aussprachewörterbuch enhält W-3LM darüberhinaus grammatikalisches Wissen in Form einer Wort-Trigram-Grammatik. Der Einsatz dieser zusätzlichen Wissensquellen erhöht den notwendigen Rechenaufwand beträchtlich. Da die erkannte Wortsequenz aber direkt in den Erkennungs- und Übersetzungsprozeß eingegliedert wird, kann der zusätzliche Aufwand vollständig einbezogen werden.

Im Testfall erhält man zu jeder unbekannten Testäußerung vier erkennerspezifische Bewertungen als Ausgabe des sprachenidentifizierenden Systems. Daher muß in der parallelen Architektur eine Klassifikationsstufe nachgeschaltet werden, die eine Entscheidung für die beste Bewertung d.h. für die korrekte Sprache trifft. Für diese Klassifikationskomponente sind zwei Gesichtspunkte zu beachten.

1. Welcher Klassifikatortyp eignet sich für die Lösung des Identifizierungsproblems am besten ?
2. Gibt es eine Transformation des Eingaberaumes, die die Aufgabe des Klassifikators effizient unterstützt ?

5.1 ML-Klassifikation

Die einfachste und naheliegenste Lösung ist die Entscheidung für diejenige Sprache, deren Erkenner den besten Score zu einem gegebenen Eingabesatz liefert (Maximum Likelihood oder ML-Klassifikator). Allerdings können die erkennerspezifischen Scores bedingt durch unterschiedliche akustische Merkmalsräume, verschiedene Parametereinstellungen am Erkenner u.v.m. in ihrem Wertebereich stark variieren. Daher sollte eine Normierung der Scores durchgeführt werden. Wir untersuchen verschiedene Normierungen und vergleichen diese mit in der Literatur angewandten Verfahren. Die Tabelle 2 zeigt für eine Auswahl von Normierungsvarianten die Identifizierungsleistung beim 4-Sprachentest mit dem ML-Klassifikator. Für das Systems W-3LM wird aufgrund der Maximum-Likelihood Entscheidung auf nicht normierten Scores stets dieselbe Sprache identifiziert, was zu dem unsinnigen Ergebnis

Normierungsart	P-3PT	W-3LM
ohne Normierung	30,6%	25,0%
Subtraktion des Mittelwertes	27,4%	35,4%
Wertebereichsnormierung	37,1%	72,5%

Table 2: Normierungen für den ML-Klassifikator

von 25% führt. Daran sieht man, daß bei der parallelen Architektur auf eine Normierung der Erkennerscores nicht verzichtet werden kann. Zissman schlug in [8] eine Normierung vor, bei der vom aktuell berechneten Score eines Erkenners der Mittelwert aller von diesem Erkenner errechneten Scores abgezogen wird. Diese Normierung führt in unseren Experimenten beim W-3LM System zu Verbesserungen. Am erfolgreichsten ist allerdings eine Normierung über den Wertebereich, bei der der aktuelle Score durch die Summe aller Ausgaben eines Erkenners auf dem gesamten Trainingsmaterial dividiert wird. Das führt zu Verbesserungen beim P-3PT System, vorallem aber beim System W-3LM. Die Leistung des ML-Klassifikators beim 4-Sprachentest sind allerdings mit 72,5% noch nicht sehr befriedigend.

Sprachen	W-3LM
4-Sprachentest	72,5%
3-Sprachentest	
Deutsch-Englisch-Spanisch	76%
Deutsch-Englisch-Japanisch	87%
Deutsch-Spanisch-Japanisch	78%
Englisch-Spanisch-Japanisch	70%
2-Sprachentest	
Deutsch - Englisch	91%
Deutsch - Spanisch	79%
Deutsch - Japanisch	99%
Englisch - Spanisch	79%
Englisch - Japanisch	89%
Spanisch - Japanisch	75%

Table 3: Testergebnisse mit ML-Klassifikator für das System W-3LM

Die Tabelle 3 zeigt die Ergebnisse aller Sprachentests bei der ML-Klassifikation für das System W-3LM . Bei Betrachtung der 2-Sprachentests fällt auf, daß die Trennung zwischen Spanisch und den übrigen Sprachen besonders schwierig ist. Auch beim 3-Sprachentest sind die Kombinationen mit

Beteiligung der spanischen Sprache schlechter. Der spanische Erkenner basiert auf einer anderen Vorverarbeitung und liefert daher Scores in einem anderen Wertebereich. Dies kann offensichtlich durch die Normierung nicht vollständig kompensiert werden. Für den deutschen, japanischen und englischen Erkenner wurde dieselbe Vorverarbeitung verwendet. Die Trennung zwischen Deutsch und Japanisch ist mit 99% ausgesprochen gut, ebenso die zwischen Deutsch und Englisch mit 91%. Letztere werden in der Literatur als schwierig zu trennendes Sprachenpaar eingeschätzt.

5.2 Neuronale Netze als Klassifikator

Neben der Frage nach einer geeigneten Normierung soll ebenfalls überprüft werden, ob die Klassifikationsaufgabe durch ein neuronales Netz besser zu lösen ist, als durch den ML-Klassifikator. Wir trainierten neuronale Netze mit unterschiedlichen Topologien und verglichen deren Leistungen mit denen des ML-Klassifikators. Die Trainings- und Testmenge der neuronalen Netze besteht aus insgesamt 1256 spontan gesprochenen Äußerungen des Korpus, je 314 pro Sprache. 70% dieser Äußerungen werden zum Training der Netze, 10% als Kreuzvalidierungsmenge eingesetzt und 20% zum Testen verwendet. Die Angaben zum ML-Klassifikator in Tabelle 2 und 3 beziehen sich auf ebendiese Testmenge.

Normierungsart	P-3PT	W-3LM
VBN+Subtraktion des Mittelwertes	69,7%	80,2%
VBN+Wertebereichsnormierung	79,8%	86,3%

Table 4: Normierungen für den NN-Klassifikator

Bei den oben genannten Normierungsvarianten sind die Ergebnisse des ML-Klassifikators gleichwertig oder besser als die des NN-Klassifikators. Wird jedoch eine Längennormierung (über die Länge der gesprochenen Äußerung) in Verbindung mit einer Vektorbetragsnormierung (VBN; Betrag des Eingabevektors = 1) durchgeführt, ergeben sich wesentliche Verbesserungen in der Identifikationsleistung des neuronalen Netzes gegenüber dem ML-Klassifikator wie in Tabelle 4 zu sehen. Der ML-Klassifikator ist bezüglich beider Normierungsverfahren invariant.

Als bester neuronaler Netztypus ergab sich in den Experimenten ist ein dreischichtiges feed-forward Netz mit 20 hidden units und einer Eingabetransformation mit Wertebereichsnormierung in Verbindung mit der Vektorbetragsnormierung. Die Tabellen 5 und 6 zeigen die Ergebnisse, die mit diesem Netz erzielt werden für beide Systeme.

Der NN-Klassifikator übertrifft die Leistungen des ML-Klassifikator in allen

Sprachen	P-3PT	W-3LM
4-Sprachentest	9,8%	86,3%
3-Sprachentest		
Deutsch-Englisch-Spanisch	66%	%
Deutsch-Englisch-Japanisch	84%	92%
Deutsch-Spanisch-Japanisch	4%	90%
Englisch-Spanisch-Japanisch	0%	4%

Table 5: 3-Sprachentest und 4-Sprachentest mit Neuronalen Netzen

vier 3-Sprachentest und im 4-Sprachentest. Die unterschiedlichen Wertebereiche der Erkenner können aber auch vom neuronalen Netz nicht vollständig kompensiert werden, wie das Absinken der Leistung beim 3-Sprachentest Deutsch-Englisch-Spanisch und Englisch-Spanisch-Japanisch zeigt.

	P-3PT				W-3LM			
Response / Stimulus	D	E	S	J	D	E	S	J
Deutsch	84%	3%	11%	2%	98%	2%	0%	0%
Englisch	11%	88%	0%	0%	2%	98%	0%	0%
Spanisch	23%	14%	61%	2%	5%	32%	60%	3%
Japanisch	3%	10%	1%	85%	2%	8%	2%	88%

Table 6: Vergleich der Systeme P-3PT und W-3LM

Die Konfusionstabelle 6 zeigt die Verwechslungen der Sprachpaare für die beiden Systeme P-3PT und W-3LM im Vergleich. Mit W-3LM erzielt man eine deutlich weniger Verwechslungen und damit bessere Identifikationsleistungen, was auf den Einsatz höherer Wissensquellen wie Aussprachewörterbuch und Grammatik zurückgeführt werden kann. Die Unterscheidung von vier Sprachen ist mit 86,3% Identifikationsrate bei W-3LM sehr gut.

6 Danksagung

Die Sammlung der multilingualen Sprachdatenbasis wird vom Bundesministerium für Bildung und Forschung (BMBF) im Rahmen des VERBMOBIL-Projekts gefördert. JANUS wird teilweise durch die Advanced Research Project Agency und das amerikanische Department of Defense gefördert.

References

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

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

[3] Y. Muthusamy, K. Berkling, T. Arai, R.A. Cole und E. Barnard: *Comparison of Approaches to Automatic Language Identification using Telephone Speech* in: Proc. Eurospeech, S. 130 7-1310, Berlin 1993.

[4] Y.K. Muthusamy, R.A. Cole und B.T. Oshika: *The OGI multi-language telephone speech corpus* in: Proc. ICSLP, Banff 1992.

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

[6] T. Schultz, I. Rogina und A. Waibel: *Experiments with LVCSR based Language Identification.* Proceedings of the SRS, S. 89-94, Baltimore 1995.

[7] T. Schultz, I. Rogina und A. Waibel: *LVCSR-based Language Identification.* in: Proc. ICASSP, S. 781-784, volume 2, Atlanta 1996.

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

[9] M.A. Zissman und 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.