# Das View4You−System: End−to−End Evaluation

F. Metze, T. Kemp; Interactive Systems Laboratories, Universität Karlsruhe (TH); {metze|kemp}@ira.uka.de

## Kurzfassung

Ein interessantes Einsatzgebiet der Spracherkennung ist die Erstellung einer Transliteration von Videodaten, also einer Verschriftung der Tonspur. Dadurch wird es möglich in diesen Daten nach bestimmten Themen zu suchen, ohne daß für jede neue Anfrage die gesamte Datenbank aufwändig von Hand durchsucht werden müsste.

Dieses Verfahren wird in dem am ISL entwickelten View4You−System [2], [10] eingesetzt, welches mittels Spracherkennung Transkripte der ARD−Tagesschau erzeugt. In diesen kann dann ein Benutzer, ebenfalls per Spracheingabe, suchen lassen, um die in Bezug auf ein bestimmtes Thema relevanten Meldungen zu finden. Diese kann er dann über eine einem Web−Browser ähnliche Schnittstelle in Bild und Ton betrachten.

In diesem Beitrag wird mittels einer End−to−end Evaluation des View4You−Systems untersucht, inwiefern die Fehler des Gesamtsystems auf die Segmentierung und Erkennung der Daten zurückgehen und welchen Einfluß die eigentliche Suche in der Datenbank hat. Durch den Einsatz eines Verfahrens zum unüberwachten Training des Spracherkenners konnte der Aufwand zur Erstellung eines vollautomatischen Systems zur Aufnahme, Archivierung und Suche in Nachrichtensendungen nahezu ohne Leistungsverlust auf etwa ein dreissigstel reduziert werden.

## 1 Das View4You−System

Das View4You−System des ISL nimmt täglich automatisch die 20:00 Uhr Ausgabe der ARD−Tagesschau auf, segmentiert die Sendung in einzelne Nachrichten, erstellt offline eine Transkription derselben und speichert Audio− und Videodaten in einer Datenbank ab, welche ständig die ca. 60 letzten Sendungen umfasst. Mittels einer Web−Browser−ähnlichen Schnittstelle oder per Spracheingabe kann ein Benutzer dann Fragen, wie zum Beispiel »Was gab es Neues auf der CeBIT?« an einen Anfrageserver stellen, welcher die in der Datenbank enthaltenen Transkripte online nach passenden Beiträgen durchsucht und diese in Bild und Ton ausgibt.

Der Benutzer kann sich also an einem Computer mittels natürlichsprachlicher Anfragen einen Überblick über die Nachrichten der vergangenen zwei Monate verschaffen, wobei die Erstellung der Datenbank vollautomatisch abläuft.

Die Aufnahme der Videodaten geschieht mit handelsüblichen Geräten, indem das von dem Satelliten Astra−1b ausgestrahlte Signal abgegriffen, digitalisiert und mittels MPEG−1 komprimiert wird. Das Audiosignal wird mit MPEG Audio Layer 2 bei einer Datenrate von 192kbit/s und 44.1kHz komprimiert.

Das View4You−System besteht aus den drei Hauptkomponenten *Segmentierer*, *Spracherkenner* und *Anfrageserver (Information Retrieval)*, welche im Folgenden beschrieben werden. **Bild 1** gibt einen Überblick über das Gesamtsystem.

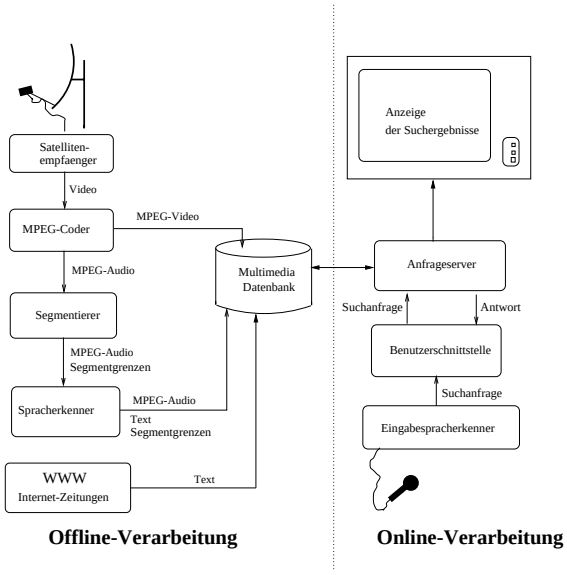


**Bild 1** Übersicht über das View4You−System

**Segmentierer**

Die aufgenommenen Sendungen müssen zuerst in einzelne Berichte segmentiert werden, da der Benutzer nicht wegen einer kurzen Meldung die gesamte Tagesschau betrachten möchte und der Spracherkenner die ungeschnittene Tagesschau (mit einer Dauer von 15 bis 20 Minuten) nicht verarbeiten kann. Jede Sendung wird daher, ausgehend von den akustischen Eigenschaften des Signals über eine Zeitskala von etwa 1s, in mehrere Segmente zerlegt.
Im Rahmen des View4You−Systems wurden vier Segmentierungsverfahren untersucht, welche nur auf Audiodaten arbeiten, die Videodaten also unberücksichtigt lassen:

- **Stilledetektion:** eine Grenze zwischen einzelnen Berichten wird eingefügt, wenn die Lautstärke unter einen bestimmten Schwellwert fällt.
- **Modellbasiert:** es werden akustische Modelle für verschiedene akustische Bedingungen (*Klassen*, in unserem Fall *Musik*, *Korrespondentenbericht*, *Studiosprecher* und *Stille*) trainiert, welche dann, ähnlich wie in der Spracherkennung, »erkannt« werden, wobei Grenzen an den Übergängen zwischen den Klasen eingefügt werden.
- **Abstandsbasiert:** zwei kurze, benachbarte Abschnitte des Audiosignals werden mittels eines Abstandsmaßes verglichen. Liegt der Unterschied über einem Schwellwert, wird eine Segmentgrenze hypothetisiert. In unserem System wurde als Abstandsmaß die Gish−Distanz [8] verwendet.
- **Hybrider Ansatz:** ein modellbasiertes Verfahren, dessen akustische Modelle für eine Klasse jeweils auf den Daten trainiert werden, die zueinander den geringsten Abstand aufweisen. Damit soll eine Kombination der Eigenschaften von modellbasierten und abstandsbasierten Verfahren erreicht werden. Dieses Verfahren wurde im Rahmen des View4You−Systems neu entwickelt [4].

Die Leistung eines Segmentierers kann daran gemessen werden, wie viele der gefundenen Segmentgrenzen tatsächlich vorhanden sind (*Precision*, *PRC*) bzw. wie viele vorhandenen Grenzen tatsächlich gefunden werden (*Recall*, *RCL*). Es gilt also:

$$PRC = \frac{Anzahl\ korrekt\ gefundener\ Grenzen}{Anzahl\ gefundener\ Grenzen}$$

$$RCL = \frac{Anzahl\ korrekt\ gefundener\ Grenzen}{Anzahl\ vorhandener\ Grenzen}$$

Im Fall des View4You−Systems wurde dabei jeweils Abweichungen von unter 1.5s noch als korrekte Zuordnung interpretiert. *Precision* und *Recall* sind nicht voneinander unabhängig. Durch die Wahl des *Arbeitspunktes* kann bestimmt werden, ob das System gute Werte für *PRC* oder *RCL* liefern soll. Die wechselseitige Abhängigkeit von *PRC* und *RCL*, die sogenannte *ROC−Kurve (Receiver−Operator−Characteristics−Kurve)*, ist für die hier untersuchten Verfahren in **Bild 2** aufgetragen.

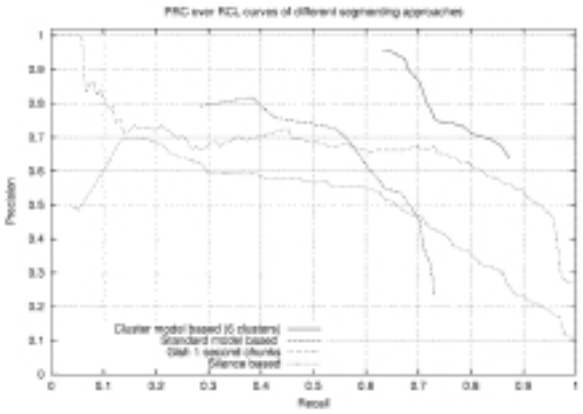


**Bild 2** ROC−Kurve verschiedener Algorithmen zur Segmentierung

Zur Bestimmung des Arbeitspunktes und zur Beurteilung des Segmentierers in einer einzigen Zahl wird häufig die sogenannte *F−Measure F* [11] herangezogen. Diese ist definiert als:

$$F = \frac{2 * PRC * RCL}{PRC + RCL}$$

F, PRC und RCL liegen also zwischen 0 und 1, wobei höhere Werte eine bessere Leistung bedeuten. Die verschiedenen Segmentierer des View4You−Systems erzielten folgende Leistungen:[1]

| Algorithmus | RCL | PRC | F−Measure |
|---|---|---|---|
| Stilledetektion | 0.62 | 0.54 | 0.58 |
| Modellbasiert | 0.56 | 0.7 | 0.62 |
| Abstandsbasiert | 0.8 | 0.63 | 0.7 |
| Hybrider Ansatz | 0.67 | 0.93 | 0.78 |

Der im Rahmen des View4You−Projektes entwickelte hybride Ansatz kombiniert die hohe Präzision des modellbasierten Ansatzes mit der hohen Recall−Rate des abstandsbasierten Ansatzes.
Eine ausführlichere Diskussion dieser Experimente findet sich in [4].

1 Optimiert wurde dabei die *F−Measure*.

**Spracherkenner**

Ein großer Teil des Aufwandes beim Einsatz von Spracherkennung besteht in der Bereitstellung der Trainingsdaten für den Erkenner. Zu jeder Aufnahme müssen sogenannte Transliterationen erstellt werden, welche zur Generierung des Sprachmodells sowie zum Training der akustischen Modelle dienen. Da der Aufwand zur manuellen Erstellung solcher Transliterationen sehr hoch ist[2] und derartige Datenbanken typischerweise einen Umfang von 10 bis 100 Stunden haben, ist ein Verfahren zu deren automatischer Erstellung von höchstem Interesse.

Eine Lösungsmöglichkeit besteht im Einsatz eines initialen Spracherkenners, dessen Ausgabe, die sogenannten »Hypothesen«, mit Konfidenzen versehen sind: dieser Erkenner wird anstelle menschlicher Experten benutzt, um die Trainingsdaten des eigentlichen Erkenners zu verschriften. Dabei »darf« eine gewisse Anzahl von Fehlern gemacht werden, welche allerdings als solche gekennzeichnet werden müssen. Auf den als »korrekt« gekennzeichneten Teilen der Hypothesen, deren Umfang ein Vielfaches der zum Training des initialen Erkenners verwendeten Daten sein wird, kann dann ein weiterer, besserer Spracherkenner trainiert werden. Der initiale Erkenner kann so auf wenig Daten trainiert werden, was die Entwicklungszeit und −kosten senkt.

Der im View4You−System zur Erzeugung der Transkripte der Tagesschau verwendete Spracherkenner wurde auf den von einem initialem Erkenner generierten, konfidenzannotierten Transkripten unüberwacht trainiert. Dabei wurden nur die Teile berücksichtigt, die eine Konfidenz von 0.5 oder besser aufwiesen. Dieser Wert wurde in einem Experiment als optimal bestimmt. Zum Vergleich wurde ein Spracherkenner auf manuell transliterierten Daten trainiert.

Das im Erkennertraining eingesetzte Konfidenzmaß *gamma* [1] ist eine a−posteriori Wahrscheinlichkeit für die Korrektheit der Worte in der besten Hypothese im Worthypothesengraph.[3] Dazu wird der Graph in ein HMM umgewandelt, so dass anschließend mittels des Forward−Backward−Algorithmus die Wahrscheinlichkeit dafür bestimmt werden kann, dass zu einer bestimmten Zeit ein bestimmter Zustand (in diesem Fall ein Wort) angenommen wird.

2 Er beträgt typischerweise 20− bis 100−fache Echtzeit und erfordert geschulte Transliterierer.

3 Dieser stellt die Ausgabe eines Spracherkenners dar.

Die Tagesschau wird zu etwa 60% von Männern moderiert. Ebenfalls 60% der Audiodaten, beispielsweise bei Korrespondentenberichten, sind mit Hintergrundgeräuschen, Musik oder Ähnlichem verunreinigt während 40% der Aufnahmen Studioqualität aufweisen. Zur Spracherkennung werden die akustischen Daten in 16kHz PCM umgerechnet. Die Vorverarbeitung besteht aus einer Berechnung der Cepstralkoeffizienten sowie deren erster und zweiter Ableitung mit anschließender LDA−Transformation und Dimensionsreduktion auf 16 Parameter pro 10ms Fenster. Die Erkenner arbeiten mit Vokaltraktlängennormierung (VTLN) sowie Sprecheradaption mittels MLLR. Das Phoneminventar besteht aus 44 Phonemen, welche aus SAMPA abgeleitet wurden. Jedes Nicht−Stille−Phonem wird als lineares HMM mit drei Zuständen modelliert. Die akustischen Modelle der finalen Erkenner bestehen aus 5.000 Triphonen, welche jeweils durch 30 Gaußverteilungen mit diagonalen Kovarianzmatrizen modelliert werden. Das Vokabular umfasst ca. 60.000 Worte, das verwendete Trigramm−Sprachmodell wurde auf 100.000.000 Worten, welche zum größten Teil aus Zeitungstexten stammen, trainiert. Die Perplexität des Sprachmodells beträgt 381 und die Out−of−Vocabulary−Rate 4.43%.

Die eingesetzten Erkenner erreichten folgende Erkennungsleistungen:

| System | Trainingsdaten | Wortfehlerrate |
|---|---|---|
| Initial | 0.5h | 32.10% |
| Unüberwachtes Training | 47.5h | 20.70% |
| Manuell erstellte Referenzen | 15.5h | 19.50% |

Durch das Konfidenzmaß wird also erreicht, dass die Hypothesen eines initialen Spracherkenners (Wortfehlerrate ca. 32%) zum Training eines neuen Spracherkenners verwendet werden können, wobei die Erkennungsleistung des neuen Spracherkenners (Wortfehlerrate ca. 21%) sehr nahe an der Leistung eines auf Transkripten trainierten Spracherkenners liegt. Da die Datensammlung vollautomatisch abläuft, konnte der Umfang der Trainingsdaten ohne Mehraufwand verdreifacht werden.
Bei unüberwachter Verwendung von 15.5h Trainingsdaten ergab sich eine Wortfehlerrate von 22.4%; dabei ist jedoch zu berücksichtigen, daß wegen der Fehler des initialen Erkenners ein Teil der Daten nicht zum Training verwendet werden können.

Der Spracherkenner des View4You−Systems wird in [3] und [10] eingehend behandelt.

**Anfrageserver/ Information Retrieval (IR)**

Zur Identifizierung der für eine Anfrage relevanten Segmente wird ein Ähnlichkeitsmaß zwischen der Anfrage und den Transliterationen der gefundenen Segmente berechnet. Dieses Verfahren findet auch bei den bekannten Internet−Suchmaschinen Verwendung. Im View4You−System wurden zwei Verfahren verglichen:

- **Okapi**: ein Vektormodell−basiertes Abstandsmaß für Suchanfragen; es vergleicht die Häufigkeit der Worte in der Suchanfrage und den Texten in der Datenbank. Dieses Maß wurde in den NIST TREC IR−Untersuchuchungen detailliert evaluiert [6].
- **Latent Semantic Indexing (LSI)**: durch Zerlegung und Dimensionsreduktion einer Term−Dokument−Matrix soll eine höhere Generalisierungsfähigkeit als bei Verwendung von Okapi erreicht werden [7].

Eine ausführlichere Beschreibung dieser Verfahren und ihrer Anwendung im View4You−System findet sich in [5]. Die Vorverarbeitung der Suchbegriffe im View4You−System besteht, wie in den meisten IR−Systemen üblich, aus Normalisierung von Sonderzeichen und Punktuation, Dekapitalisierung, Eliminierung von Stoppworten und Zurückführen auf die Stammform.

Die Leistung eines IR−Systems bemisst sich, ähnlich wie bei der Segmentierung, nach dem Prozentsatz der wiedergefunden Beiträge (*Recall, RCL*) und den Anteil relevanter Beiträge in der Antwort (*Precision, PRC*). Diese Werte können ebenfalls in einer ROC−Kurve aufgetragen werden, um das Verhalten eines Systems vollständig zu beschrieben. In der Literatur werden häufig die *Average Precision* (*AveP*) bzw. *R−Precision* (*R−PRC*) angegeben, welche eine Bewertung des Systems durch eine einzige Zahl zu erreichen versuchen. Dabei gilt:

$$AveP = \frac{1}{11} \sum_{x \in \{0, 0.1, \ldots, 1.0\}} PRC\left(RCL = x\right)$$

*AveP* ist also eine mittlere *PRC*, welche an 11 äquidistant gelegenen Werten von *RCL* ermittelt wird.
Als *R−PRC* ist derjenige Wert von *PRC* definiert, der erreicht wird, wenn für jede Anfrage die Systemantwort auf die *R* relevantesten Segmente eingeschränkt wird. *R* wird so bestimmt, dass die Länge dieser Segmente der Gesamtlänge aller in der Referenz relevanten Segmente entspricht. *R−PRC* wird über alle Anfragen gemittelt [9].
Diese Maße wurden auch für die End−to−End Evaluation eingesetzt. Die Leistung der IR−Komponente ist aus Vergleichsgründen im folgenden Kapitel zusammen mit der End−to−End−Evaluation dargestellt; sie ergibt sich aus der Systemleistung bei Verwendung von manueller Segmentierung und Transkriptionen.

# 2 End−to−end Evaluation

Die Referenzausgabe des Gesamtsystems besteht aus den relevanten[4] manuell segmentierten einzelnen Nachrichtenblöcken der in der Datenbank enthaltenen Sendungen. Die für die End−to−End Evaluation verwendete Datenbank enthielt 65 Aufnahmen der Tagesschau mit 3165 manuell bestimmten Segmenten. Die Evaluation wurde mit 10 Anfragen, die typische Anfragen naiver Benutzer repräsentieren, durchgeführt. Sie umfassen sowohl regelmäßige Ereignisse (»Ich möchte die Lottozahlen wissen.«), als auch Ereignisse, die nur für kürzere Zeit von Interesse waren (»Wird Helmut Kohl noch einmal kandidieren?«).

Die End−to−End−Evaluation des View4You−Systems soll beantworten, wie groß der Einfluss der drei Hauptkomponenten *Spracherkenner*, *Segmentierer* und *Anfrageserver / Information Retrieval* auf die Leistung des Gesamtsystems ist.

| ***Segmentierung, Textbasis*** | ***Okapi (AveP)*** | ***LSI (AveP)*** |
|---|---|---|
| Manuell, Transkriptionen | 0.45 | 0.59 |
| Manuell, Hypothesen | 0.43 (−4.2%) | 0.57 (−3.5%) |
| Hybrid, Hypothesen | 0.36 (−20.5%) | 0.51 (−14.5%) |
| Modellbasiert, Hypothesen | 0.28 (−38.1%) | 0.48 (−19.6%) |

4 Die Relevanz der Meldungen wurde dabei manuell durch einen Benutzer festgestellt, welcher die gesamte Datenbank kannte.

| *Segmentierung, Textbasis* | *Okapi (R−PRC)* | *LSI (R−PRC)* |
|---|---|---|
| Manuell, Transkriptionen | 0.45 | 0.58 |
| Manuell, Hypothesen | 0.43 (−4.9%) | 0.56 (−4.5%) |
| Hybrid, Hypothesen | 0.39 (−12.3%) | 0.49 (−16.7%) |
| Modellbasiert, Hypothesen | 0.30 (−33.6%) | 0.47 (−18.9%) |

Die Ergebnisse der End−to−End−Evaluation sind in den vorangestellten Tabellen anhand der Größen *AveP* und *R−PRC* dargestellt. Die Werte in Klammern geben die relative Verschlechterung gegenüber der Referenzausgabe, welche auf manueller Segmentierung und Verwendung von Transkriptionen beruht, an.

Man erkennt, dass selbst die Verwendung eines unüberwacht trainierten Spracherkenners zur Erstellung der Transliterationen, welche als Basis des Information−Retrieval dienen, nur zu einer geringfügigen Verschlechterung (kleiner 5%) der Gesamtleistung, ausgedrückt in *Average Precision* bzw. *R−Precision* führt.
Einen großen Einfluss auf die Gesamtleistung hat hingegen das verwendete Segmentierungsverfahren. Der neu entwickelte hybride Ansatz stellt eine deutliche Verbeserung gegenüber bisherigen Verfahren dar. Der durch eine falsche Segmentierung bedingte Leistungsabfall ist dennoch etwa drei mal so stark wie der vom Einsatz eines Spracherkenners her rührende. Auffällig ist der starke Abfall an IR−Leistung beim Einsatz des modellbasierten Verfahrens mit *Okapi*, welcher bei Verwendung von *LSI* nicht auftritt.
Selbst unter optimalen Bedingungen (Manuelle Segmentierung, Verwendung von Transkripten) werden vom Anfrageserver nur maximal 60% der relevanten Beiträge identifiziert. Die kritische Komponente des View4You−Systems und vergleichbarer Projekte ist also der Information−Retrieval Teil.

# 3 Zusammenfassung

Das View4You−System ist ein vollautomatisches System zur Indexierung und Abfrage von Fernsehnachrichten. Dieser Beitrag präsentierte die Ergebnisse einer End−to−End−Evaluation des Gesamtsystems mit 10 Anfragen an eine Datenbank, welche 65 Sendungen der »Tagesschau« enthält.

Die Erzeugung der Transkripte, welche in die Datenbank aufgenommen werden, geschieht durch einen Spracherkenner, welcher ohne Verwendung von Transkripten unüberwacht trainiert wurde, wodurch kein signifikanter Verlust an Systemleistung entsteht. Der wesentlich stärkere Leistungsverlust durch die suboptimale Segmentierung der Sendungen in einzelne Nachrichten konnte durch einen neuentwickelten hybriden Segmentierer deutlich reduziert werden. In der End−to−End−Evaluation wurde *LSI* (*Latent Semantic Indexing*) mit klassischen Verfahren des Information Retrieval (*Okapi*) verglichen. Die auf den Erkennerhypothesen und automatischer Segmentierung unter Verwendung von *LSI* erreichte System−Leistung (*R−PRC* = 0.49) ist besser als die mit *Okapi* und manueller Transkription und Segmentierung erreichte (*R−PRC* = 0.45). Gleichzeitig zeigt *LSI* auch eine wesentlich geringere Abhängigkeit von der Art der verwendeten Segmentierung, so dass sich dieses Verfahren für den Einsatz in einem System zur Abfrage von automatisch aufgenommen Nachrichten deutlich besser eignet.

Gegenwärtig hat also die Indexerstellung, bzw. die Segmentierung unformatierter Nachrichten, den größten Einfluß auf die Leistung eines automatischen »News−on−demand« Systems. Durch Latent Semantic Indexing konnte die Systemleistung stark verbessert werden, allerdings benötigt dieses Verfahren zur Indexerstellung sehr viel Rechenzeit. Weitere Untersuchungen sollten daher hinsichtlich der Skalierbarkeit vor allem der IR−Leistung mit zunehmender Datenbankgröße erfolgen.

# 4 Literatur


[1] T. Kemp, T. Schaaf: *Estimating confidence using word lattices*, in Proc. EUROSPEECH−97, Vol. 2; Rhodos, Griechenland, 1997

[2] T. Kemp, A. Waibel, et al.: *Automatische Erstellung einer Video−Datenbank: das View4You−System*, in Proc. 4th Conference on Natural Language Processing, Vol. 1; Bonn, Deutschland, 1998

[3] T. Kemp, A. Waibel: *Unsupervised training of a speech recognizer: recent experiments*, in Proc. EUROSPEECH−99, Vol. 6; Budapest, Ungarn, 1999

[4] T. Kemp, M. Schmidt, M. Westphal und A. Waibel; *Strategies for automatic segmentation of audio data*, in Proc. ICASSP−2000; Istanbul, Türkei, 2000

[5] T. Kemp, M. Weber und A. Waibel: *Evaluating different information retrieval algorithms on real−world data*, in Proc. ICSLP−2000; Peking, VR China, 2000

[6] M. M. Beaulieu, M. Gatford, X. Huang, S. E. Robertson, S. Walker und P. Williams, *Okapi at TREC−5;* in Proc. 5th Text Retrieval Conference; Gaithersburg (MD), USA, 1997

[7] S. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas und R. A. Harshman, *Indexing by Latent Semantic Indexing;* Journal of the Society for Information Science, 1990

[8] H. Gish, M. H. Siu und R. Rohlicek: *Segregation of speakers for speech recognition and speaker identification*, in Proc. ICASSP−91; Toronto, Kanada, 1991

[9] R. Baeza−Yates, B. Ribeiro−Neto, *Modern Information Retrieval*, ACM Press, New York, ISBN 0−201−39829−X, Addison−Wesley, 1999

[10] T. Kemp: *Ein automatisches Indexierungssystem für Fernsehnachrichten sendungen*, Dissertation, Universität Karlsruhe, ILKD; Karlsruhe, Deutschland, 2000

[11] C. J. Van Rijsbergen, *Information Retrieval*, Butterworth, London, 1979