Hochschule für Angewandte Wissenschaften Hamburg
Hamburg University of Applied Sciences

Ausarbeitung Anwendungen 2 - SoSe 2009

Kristoffer Witt

Automatische Transkription von Radiowerbespots
Stand der Forschung

Fakultät Technik und Informatik
Department Informatik

Faculty of Engineering and Computer Science
Department of Computer Science

Inhaltsverzeichnis

1 Einführung in das Themengebiet **3**

2 Stand der Forschung **3**
2.1 Forschungshistorie 4
2.2 Aktuelle Herausforderungen 5
2.3 Vergleichbare Arbeiten 6
2.4 Analyse 10

3 Abgrenzung eigener Arbeitsziele **10**

4 Zusammenfassung **12**

Literatur **12**

Kurzzusammenfassung

Im Folgenden wird der Forschungsstand auf dem Gebiet der Spracherkennung erörtert und hinsichtlich der Nutzung für die Transkription von Radiowerbespots analysiert.

Abstract

This report presents the current state of the art of speech recognition research in regards to automatic transcription of radio advertisements.

1 Einführung in das Themengebiet

Die Überführung von gesprochener Sprache in eine textuelle Repräsentation, war und ist schon seit über 50 Jahren (O'Shaughnessy, 2008, Seite 2966ff) ein aktives Forschungsgebiet von Wissenschaftlern auf der ganzen Welt. Meilensteine, wie der Einsatz von Hidden Markov Modellen (HMM) und die Sprachkodierung mit Mel-Frequenz Cepstren haben die Qualität der automatischen Spracherkennung (engl. Automatic Speech Recognition, ASR) in hohem Maße verbessert (O'Shaughnessy, 2008, Seite 2967)(Baker u. a., 2009, Seite 75). Dabei hat der voranschreitende Technologische Fortschritt, z.B. Erhöhung der Speicher- und Rechenkapazitäten, erst die Erfassung und Zurverfügungstellung von immer umfassenderen Audio- und Sprachmodellen, die für die ASR von essentieller Wichtigkeit sind (Rabiner, 2003, Seite 1495) ermöglicht.

Im Folgenden Artikel wird der aktuelle Stand der Forschung im Bereich ASR hinsichtlich der Nutzung für die automatische Transkription von Rundfunk Werbespots analysiert. Für die Motivation dieser Arbeit sei auf Witt (2008) verwiesen. Einleitend wird der aktive Forschungsbereich ASR, seine Bestandteile und eine Historie derer kurz erläutert. Nach der Vorstellung thematisch ähnlicher Veröffentlichungen (siehe Bereich Verwandte Arbeiten), erfolgt eine Einordnung der geplanten eigenen Arbeit bezüglich der aktuellen Forschungsgebiete/-zweige.

Abschließend wird in einer Zusammenfassung kurz dargestellt welche Schlüsse die Analyse der gegenwärtigen Forschung gebracht hat.

2 Stand der Forschung

Neben einem kurzen Abriss über die Entwicklung der letzten 50 Jahre werden in diesem Abschnitt grundlegende Begriffe erläutert, die für das Verständnis von besonderer Wichtigkeit sind.

Advance	Date	Impact
Linear predictive coding	1969	Automatic, simple speech compression
Dynamic time warping	1970s	Reduces search while allowing temporal flexibility
Hidden Markov models	1975	Treat both temporal and spectral variation statistically
Mel-frequency cepstrum	1980	Improved auditory-based speech compression
Language models	1980s	Including language redundancy improves ASR accuracy
Neural networks	1980s	Excellent static nonlinear classifier
Kernel-based classifiers	1998	Better discriminative training
Dynamic Bayesian networks	1999	More general statistical networks

(O'Shaughnessy, 2008, s. 2967)

Abbildung 1: Meilensteine der Automatischen Spracherkennung

2.1 Forschungshistorie

Obwohl bereits seit mehr als 50 Jahren auf dem Gebiet der Computer gestützten Erkennung von Sprache geforscht wird, ist man von der Leistung die ein Mensch vollbringt noch weit entfernt. Aus Abbildung 1 (O'Shaughnessy, 2008, Seite. 2967) können die Meilensteine entnommen werden, die laut O'Shaugnessy den größten Einfluss auf den Fortschritt der ASR hatten.

Linear Predictive Coding Ein Verfahren zur Codierung bzw. Kompression von Sprachsignalen. Ermöglichte es erstmals, unter den damalig eingeschränkten Bedingungen, Sprache zu verarbeiten (O'Shaughnessy, 2008, Seite 2967).

Dynamic Time Warping Dient der Kompensation von temporalen Variationen bei der Aussprache von Phonemen/Phonen.

Hidden Markov models Erstellen Wahrscheinlichkeitsmodelle für die Erkennung von Sprache. Ermöglicht eine starke Reduktion der zu vergleichenden Strukturen, im Vergleich zur Template basierten Spracherkennung.

Mel-frequency cepstrum Eine logarithmische Einteilung des Frequenzraumes, die sich an der menschlichen Perzeption orientiert. Siehe Article 569587 für nähere Informationen.

Language Models Statistische Erfassung der Wort-Häufigkeiten einer Sprache. Ermöglicht es Wahrscheinlichkeiten von bestimmten Wortkombinationen zu berechnen und damit die Erkennung zu verbessern.

Neural Networks Versuch der Adaption der Funktionsweise des menschlichen Gehirns. Ermöglicht eine Klassifizierung anhand im Vorfeld gelernter Parameter.

Kernel Based Classifiers Automatische Klassifizierung von Daten ähnlich der Funktionsweise von Neuralen Netzwerken.

Dynamic Bayesian Networks Statistische Auswertung von Sprache zur Verbesserung der Erkennung. Siehe Ghahramani (1998) für zusätzliche Informationen.

Die oben beschriebenen Techniken und Technologien bilden, obwohl größtenteils schon vergleichsweise alt, immer noch die Basis für aktuelle Spracherkenner. Insbesondere die Repräsentation von Sprach- und Audio-Modellen als Hidden-Markov-Modelle sowie die Mel-Frequenz-Cepstren. Diese sind das am häufigsten verwendete Merkmal bei der Audiodatenmodellierung (Fang u. a., 2001), da sie von Stimmeneigenheiten sehr gut abstrahieren und damit die Phonem-Erkennung verbessern.

Neben den beschriebenen Verfahren und Techniken müssen auch einzelne Forschungsgruppen erwähnt werden. Deren Engagement erst den heutigen Wissensstand ermöglichte. Aufgrund der hohen Komplexität des Spracherkennungsproblems ist die Synthese von für die Forschung geeigneten Werkzeugen nur Spezialisten vorbehalten. Durch die Zurverfügungstellung dieser Werkzeuge wurde weitere Forschung auf unterschiedlichen Gebieten erst ermöglicht (Baker u. a., 2009, Seite 75). Zum Beispiel der Hidden Markov Toolkit(HTK) der Cambridge University und das SPHINX Projekt der Carnegie Mellon University. Finanziert wurde ein Großteil der Projekte von der U.S. Department of Defense Advances Projects Research Agency (Baker u. a., 2009, Seite 75).

Weitere interessante Ressourcen können (Nguyen, 2009) entnommen werden.

2.2 Aktuelle Herausforderungen

Baker et al. beschreiben die sogenannten „Grand Challenges“ der Spracherkennung als

> „Ambitious but achievable three-to fiveyear research program initiatives that will significantly advance the state of the art in speech recognition and understanding“

(Baker u. a., 2009, Seite 76). Also als ehrgeizige Projekte, die innerhalb von drei bis fünf Jahren den „state of the art“ der Sprachverarbeitung signifikant voranbringen können.

Nachfolgend werden die für den Kontext dieser Arbeit interessanten „Grand Challenges“ aufgelistet und erläutert.

Everyday Audio und Selbstlernende Adaptive Erkenner

Abstrakt beschrieben sind aktuelle Spracherkenner stark spezialisierte Experten. Das bedeutet sie sind meist auf die Erkennung eines Sprechers in einer bestimmten Umgebung mit bekannten Parametern trainiert. Sobald sich diese Variablen allerdings ändern, verschlechtert sich die

Rate der korrekt erkannten Wortsequenzen rapide. Dies trifft besonders auf „normale“ Konversationen, also Dialoge mit zwei oder mehr Teilnehmern, zu. Im Gegensatz zu vorgelesenem Text gibt es bei dieser Art Konversation keine lineare Struktur. Sprecher unterbrechen sich gegenseitig, Füllwörter wie „äh“ werden geäußert oder andere Störgeräusche finden ihren Weg in die Aufnahme. Der Bereich „Everyday Audio“ beschäftigt sich mit der Korrektur dieses Missstands. Dabei wird insbesondere versucht, adaptive Spracherkenner zu entwickeln, die auf sich ändernde Parameter schnell reagieren können. Dies ist für verschiedene Sprecher bereits heute in gewissem Maße möglich, siehe (Huang und Lee, 1991).

Detection of rare, key events

Besondere Schwierigkeiten bereiten Spracherkennern die sogenannten Out-Of-Vocabulary (OOV) Wörter. Das sind Wörter, die nicht im trainierten Sprach-/Audiocorpus vorkommen. Die Erkennung, dass es sich bei dem zu erkennen Wort um ein unbekanntes Wort handelt, bezeichnet man auch als „rare-“ oder „key-Event“. Dieses Forschungsgebiet versucht Spracherkennung so zu verbessern, dass unbekannte Wörter auch als solche klassifiziert werden und nicht als ähnlich klingende im Sprachkorpus vorhandene Wörter.

Verbesserung der ASR-Infrastruktur

Unter den Begriff der Infrastruktur fällt die Erstellung von qualitativ hochwertigen Sprach- und Audiokorpora. Wie bereits beschrieben, ist das Fundament eines Spracherkenners seine Trainingsdaten. Je besser diese auf den späteren Einsatzzweck abgestimmt sind, d.h. je genauer sie die Umgebungsbedingungen und den Wortumfang repräsentieren, desto besser ist auch das zu erwartende Ergebnis. Somit ist ein für die Voranbringung der Forschung besonders wichtiger Punkt die Erzeugung von unterschiedlichen und umfangreiche Korpora.

2.3 Vergleichbare Arbeiten

Der Folgende Abschnitt beschreibt zwei vom Grundaufbau diese Arbeit ähnliche Arbeiten. D.h. sie befassen sich mit Radio- bzw. Radioähnlichem Sprachmaterial. Die beiden vorgestellten Arbeiten bilden nur einen kleinen Teil der veröffentlichten Literatur. Ihre Ergebnisse spiegeln den Konsens der weiteren verfügbaren Arbeiten wider, vgl. siehe Kokaram u. a. (2006), Chelba u. a. (2008), Chen u. a. (2006)

Automatic transcription of general audio data: Effect of environment segmentation on phonetic recognition (Spina und Zue, 1996)

Die Autoren Michelle S. Spina und Victor W. Zue gehören der Spoken Language Systems Group des Laboratory for Computer Science der Massachusetts Institute of Technology an. Als das Hauptziel Ihrer Arbeit beschreiben Sie das Finden einer optimalen Trainingsstrategie für die Erkennung von generischem Radio und TV Sprachmaterial, der sogenannten „general audio data“(GAD)(Spina und Zue, 1996, Seite 1). Begründet damit, dass durch die immer größere Verbreitung von GAD-Daten, eine inhaltliche Erfassung mehr und mehr an Wichtigkeit gewinnt. Insbesondere hinsichtlich Anwendungsbereichen wie zum Beispiel Durchsuchung und Indexierung (vgl. (Ranjan u. a., 2006)).

Die Versuchsdaten bestehen aus sechs einstündigen Aufnahmen des Radioprogramms „Morning Edition“ des National Public Radio. Sie besteht aus Reportagen, Moderationen, Interviews und Musikpassagen. Die Aufnahmen wurde im Zeitraum November 1996 bis Januar 1997 erstellt.

Diese Daten wurden zunächst einmal in sieben Kategorien segmentiert:

1. Saubere Sprache („"clean speech“), Breitband, 8 kHz, Moderatoren und Reporter im Studio aufgenommen.
2. Musik Sprache („music speech“), Sprache mit Musik im Hintergrund.
3. Verrauschte Sprache („noisy speech“), Sprache mit Hintergrundlärm.
4. Vor-Ort-Sprache („field speech“), Telefon 4 kHz, Sprache von Reportern Vor-Ort.
5. Musik („music“)
6. Stille („silence“)
7. Müll („garbage“)

Segmentiert wurde automatisch anhand verschiedener Kriterien die Spina und Zue (1996) entnommen werden können. Auf die Segmentierung folgend, wurde für die ersten vier Kategorien eine Spracherkennung durchgeführt. Der dabei eingesetzte, auf den TIMIT-Daten (Garofolo u. a., 1993) trainierte, Erkenner ist eine Eigenentwicklung des MIT mit Namen SSUMMIT"(Zue u. a., 1989).

Es wurden zwei verschiedene Herangehensweisen getestet. Zum einen der Einsatz eines Erkenners für alle vier Klassen. Zum zweiten für jedes Sprachsegment jeweils ein Erkenner. Das Experiment des ersten Typs wurde weiter unterteilt in unterschiedliche Trainingsarten. Das bedeutet, der einzelne Erkenner wurde zum einen mit Sprachmaterial aus allen Klassen und zum anderen nur mit sauberer Sprache trainiert. Um das Ergebnis zu verbessern wurde zusätzlich „cepstral mean normalization“(Liu u. a., 1993) auf Trainings- und Testdaten angewendet.

Training Data	Testing Data				
	Clean Speech	Music Speech	Noisy Speech	Field Speech	Over All
Multi-Style	42.3	51.8	51.8	65.1	48.2
With CMN	40.4	50.2	50.0	55.5	45.8
Clean Speech	39.8	58.2	50.9	63.9	47.7
With CMN	38.2	53.4	49.1	57.3	45.2

Abbildung 2: Wort-Fehler-Rate der einzelnen Modi

Der Vergleich der Resultate lieferte einen Vorteil zu Gunsten des Einzelerkenners. Dies ließ sich auf die manuelle Einteilung der Klassen zurückführen. Es wurde subjektiv, manuell, ein Großteil der Sprache als „noisy“ eingestuft obwohl nur geringe Hintergrundgeräusche wahrnehmbar waren. Dadurch erhöhte sich die Wort-Fehler-Rate für diese Klasse um 20%. Kompensiert durch den Einsatz eines automatischen Klassifizierers, der unsensibler gegen Hintergrundrauschen konfiguriert wurde, wurde das beste Ergebnis mit den multiplen Erkennern erreicht. Allerdings betrug der Unterschied nur 1,8% der Wort-Fehler-Rate[1]. Eine Übersicht über die Wort-Fehler-Raten liefert Tabelle 2 aus Spina und Zue (1996).

Automatic Multimedia Indexing: Combining audio, speech, and visual information to index broadcast news

Im Unterschied zu der erstgenannten Arbeit befasst sich das Paper von Ohtsuki, Bessho, Matsuo Matsunaga und Hayashi nicht ausschließlich mit Audio-Daten. Es werden zusätzlich etwaig verfügbare Videodaten zur Auswertung hinzugezogen. Es wird ein System vorgestellt, das für die Indizierung von Nachrichten anhand extrahierter „Handlungsstränge“ eingesetzt werden soll. Diese Handlungsstränge werden anhand von Sprach-, Audio- und Videodaten identifiziert. Beispielsweise Berichte über ein bestimmtes Thema, wie die Fußballweltmeisterschaft. Das Paper befasst sich zusätzlich auch mit der Geschwindigkeit der Erkennung, gemessen als Real Time Factor (RTF). Der Real Time Factor bezeichnet das Verhältnis von für die Verarbeitung benötigter Zeit zur Echtzeit in der die Sprache geäußert wurde.

Als Testdaten wurden 12 verschiedene 5-30 Minuten dauernde Nachrichten-Programme, teils mit Werbung teils ohne, verwendet. Die Grundelemente des Systems können Abbildung 3

[1] Für dieses Ergebnis wurden die Vor-Ort-Sprache Band gefiltert zwischen 133Hz und 4kHz.

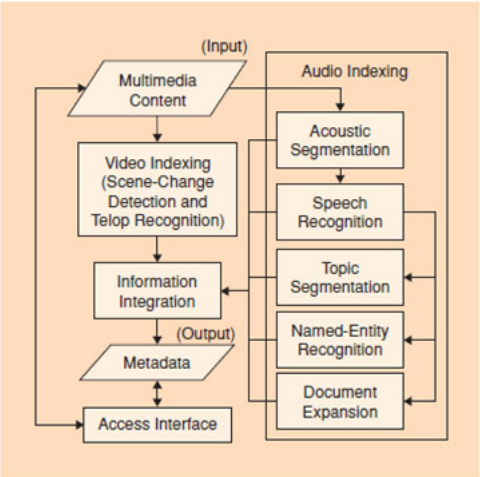

Abbildung 3: Komponenten des Gesamtsystems, aus (Ohtsuki u. a., 2006, Seite 70)

entnommen werden. Für diese Arbeit von Interesse sind hierbei die Bereiche „Acoustic Segmentation“ und „Speech Recognition“.

Acoustic Segmentation Extrahiert anhand von bestimmten akustischen Eigenschaften die Typen der Audiodaten. Aufgeteilt in Sprache, Musik, Rauschen und Stille. Nur die Sprachsegmente werden an die Spracherkennung weitergeleitet.

Speech Recognition Das Spracherkennungsmodul führt für die erkannten Sprachsegmente eine „Large Vocabulary Continuous Speech Recognition (LVCSR)“ durch. Für diesen Zweck kommt die vom Nippon Telegraph and Telphone Laboratory entwickelte Spracherkennung „VoiceRex“ zum Einsatz.

Für die Evaluation der Spracherkennung wurden verschiedene Audio-Modelle verwendet.

Single Ein geschlechtsunabhängig trainiertes Modell (GI, gender independent)

Parallel Verschiedene unterschiedliche Modelle werden angewendet, das Modell mit der höchsten Wahrscheinlichkeit wird eingesetzt.

Select Ein kurzer Teil am Anfang der Testdaten wird ausgewertet und mit den verfügbaren Modellen abgeglichen. Das Modell mit den besten Ergebnissen wird dann verwendet.

Hinsichtlich der Geschwindigkeit der Ansätze wurde festgestellt, dass die Parallele Auswertung aller verfügbaren Modelle am längsten benötigt, gefolgt von den Select- und Single-Modi. Dabei lieferten alle drei Modi qualitativ vergleichbare Ergebnisse. Am Besten hierbei schnitt die Parallele Auswertung ab (sie benötigte allerdings die dreifache Zeit verglichen mit dem Single-Ansatz).

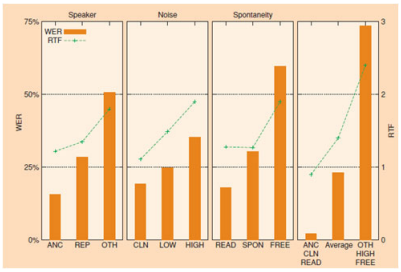

Abbildung 4: Resultate der Tests, WER - Word Error Rate, RTF - Real Time Factor. (Ohtsuki u. a., 2006, Seite 70)

Wie sich aus den Ergebnissen, dargestellt in Abbildung 4[2], entnehmen lässt, ist die Qualität der Erkennung stark korreliert mit Umgebungs- und Sprachbedingungen. Je regelmäßiger die Sprache ist, zum Beispiel vorgelesen vom Moderator, desto besser das Ergebnis. Tendiert die aufgenommene Sprache mehr zu spontaner Konversation, wie ein Interview mit einem Vor-Ort-Reporter, nimmt die Erkennungsrate stark ab. Ähnlich stellt sich das Ergebnis für den Rauschanteil da, je klarer die Sprache und somit geringerer Rauschanteil, desto besser das Ergebnis.

2.4 Analyse

Analysiert man die Resultate der vorgestellten Arbeiten, lassen sich gut verschiedene Punkte identifizieren, die für eine fehlerarme Spracherkennung von Sprachdaten essentiell sind. Um möglichst gute Ergebnisse zu erhalten ist es wichtig, speziell trainierte Modelle zu verwenden. Das bedeutet optimal wäre es, für jede mögliche Aufnahmesituation und jeden möglichen Sprachkontext ein Modell zu trainieren. Da die Anzahl der Modelle allerdings exponentiell ansteigt ist dies nahezu unmöglich. Eine Lösung für dieses Problem ist die Klassifizierung beziehungsweise Segmentierung der Daten. Die Schwierigkeit besteht darin, einen Mittelweg zwischen hoher Anzahl Modelle und geringer Klassenzahl zu finden.

3 Abgrenzung eigener Arbeitsziele

Verglichen mit den vorgestellten Arbeiten sind die angestrebten Ziele dieser Arbeit ähnlich. Die Bestimmung/Segmentierung von Sprachsegmenten in einem Audiodatenstrom mit anschließender Transformation in textuelle Form. Allein der Unterschiedliche Typ der Daten, also Ra-

[2] Abkürzungen: Moderator, Anchor (ANC), Reporter (REP), Andere, Other(OTH), Rauschfrei, Clean(CLN), Spontan, Spontaneous(SPON)

diowerbespots, unterscheidet das Ziel von vielen bereits veröffentlichten Arbeiten. Diese, auf den ersten Blick, kleine Differenz jedoch verändert das Spracherkennungsproblem grundlegend, so dass mit den vorhandenen/evaluierten Methoden und Modellen ein zufriedenstellendes Ergebnis wohl nicht erreichbar ist.

Zwecks Verdeutlichung der Unterschiede werden zunächst einmal die Eigenschaften der Transkriptionsdaten erläutert.

Eigenschaften von Radiospots

Anteil von deutlicher Sprache relativ hoch Da es sich um Werbebotschaften handelt, das bedeutet es wird ein Produkt vorgestellt das konsumiert werden soll, sollte die verwendete Sprache relativ deutlich sein. Ausnahmen bilden hierbei Spots bei denen über die Art der Sprache eine Botschaft vermittelt werden soll. Als Beispiel sei ein hypothetischer Sport einer lokalen Brauerei genannt, die versuchen könnte über die Verwendung eines starken Sprachakzents lokale Kunden zu erreichen.

Hohe Rauschvarianz Da Menschen relativ gut Musik, Sprache und Effekte auseinanderhalten können, sind etwaige Hintergrundgeräusche, Jingles oder ähnliches der Werbebotschaft nicht abträglich. Daher finden Sie relativ häufig Ihren Einsatz zur Auflockerung von Spots. Da es jedoch kein Grundschema gibt, ist eine Vorhersage des Rauschens und damit eine Festlegung auf bestimmte Modelle, nahezu unmöglich.

Anteil von OOV Wörtern Für die Verwendung des Transkripts des Spots als Unterstützung der Produktionsmitarbeiter (siehe Witt (2008)) ist es besonders wichtig das der Produkt bzw. der Markenname richtig erkannt wird. Da dies allerdings Eigennamen sind die in den heutzutage verfügbaren Sprach- und Audiomodellen mit hoher Wahrscheinlichkeit nicht vorhanden sind, gelten sie als OOV-Wörter.

Aus diesen Eigenschaften lassen sich nun folgende Schlüsse ableiten:

- Es muss geprüft werden, inwiefern vorhandene Audio- und Sprachmodelle auf die in Radiospots geäußerte Sprache passen.
- Die Sprache muss klassifiziert werden um eine möglichst geringe Anzahl Modelle für die Erkennung zu erhalten.
- Mit hoher Wahrscheinlichkeit müssen neue Modelle erzeugt, bzw. vorhandene Modelle um die Produkt- und Markenbezeichnungen ergänzt werden. Um die Anzahl der wichtigen OOV-Wörter zu minimieren.

  Zu diesem Zweck wäre es möglich, aus vorhandenen TV Werbespots Audiomodelle zu extrahieren und diese mit OCR-Daten aus Printanzeigen abzugleichen. Dabei bleibt zu klären, inwiefern eine Korrelation zwischen der Sprache der einzelnen Werbemedien besteht.

4 Zusammenfassung

Es bleibt festzustellen, dass obwohl das Forschungsgebiet schon über ein halbes Jahrhundert alt ist, die Erkennung von generischer, untrainierter, Sprache noch immer nur partiell möglich ist. Durch die hohe Varianz die in Radiospots vorherrscht, also unterschiedliche Umgebungen, Sprecher, Rauschen, erschwert ist ein Training von Spracherkennern für die fehlerfreie Transkription relativ unmöglich. Der Erfolg der angestrebten Arbeit hängt daher im großen Maße davon ab, inwiefern eine Erkennung von Marken und Produktnamen möglich ist. Da diese für die beschriebenen Einsatzziele Witt (2008) von hoher Bedeutung sind.

Die Ergebnisse der in Vielzahl vorhandenen ähnlichen Arbeiten ermöglichen eine stetige Verbesserung der Qualität. Für welchen Zweck diese zum aktuellen Zeitpunkt ausreicht sollte in Versuchen evaluiert werden.

Weiterhin soll nicht unerwähnt bleiben, dass durch die sich ständig vergrößernde Anzahl an Audiodaten, im Internet, im TV oder auch im Radio, die Bedeutung von Spracherkennung für die Indizierung analog zunimmt. Daher wird der Forschungsbereich Spracherkennung auch weiterhin so aktiv bleiben, wie er es heute ist. Dementsprechend werden auch in der nächste Zeit Methoden und Techniken entwickelt werden, die die Fehlerrate weiter verkleinern und die Performanz und damit die Einsatzmöglichkeiten, vervielfachen, siehe (O'Shaughnessy, 2008). Bis ein Computer allerdings die perzeptiven Fähigkeiten eines Menschen nachbilden kann, bleibt es noch ein weiter Weg (Scharenborg, 2007, Seiten 336 ff).

Literatur

[Androutsos u. a. 2006] ANDROUTSOS, D. ; GUAN, Ling ; VENETSANOPOULOS, A.N.: Semantic retrieval of multimedia [from the Guest Editors]. In: *Signal Processing Magazine, IEEE* 23 (2006), Mar, Nr. 2, S. 14–16. – ISSN 1053-5888

[Baker u. a. 2009] BAKER, J. ; DENG, Li ; GLASS, J. ; KHUDANPUR, S. ; LEE, Chin hui ; MORGAN, N. ; O'SHAUGHNESSY, D.: Developments and directions in speech recognition and understanding, part 1. In: *Signal Processing Magazine, IEEE* 26 (2009), May, Nr. 3, S. 75–80. – ISSN 1053-5888

[Chelba u. a. 2008] CHELBA, C. ; HAZEN, T.J. ; SARACLAR, M.: Retrieval and browsing of spoken content. In: *Signal Processing Magazine, IEEE* 25 (2008), May, Nr. 3, S. 39–49. – ISSN 1053-5888

[Chen u. a. 2006] CHEN, Ken ; HASEGAWA-JOHNSON, Mark ; COHEN, Aaron ; BORYS, Sarah ; KIM, Sung-Suk ; COLE, Jennifer ; CHOI, Jeung-Yoon: Prosody dependent speech

recognition on radio news corpus of American English. In: *IEEE Transactions on Audio, Speech & Language Processing* 14 (2006), Nr. 1, S. 232–245

[Fang u. a. 2001] FANG, Zheng ; GUOLIANG, Zhang ; ZHANJIANG, Song: Comparison of different implementations of MFCC. In: *J. Comput. Sci. Technol.* 16 (2001), Nr. 6, S. 582–589. – ISSN 1000-9000

[Garofolo u. a. 1993] GAROFOLO, J. S. ; LAMEL, L. F. ; FISHER, W. M. ; FISCUS, J. G. ; PALLETT, D. S. ; DAHLGREN, N. L.: *DARPA TIMIT Acoustic Phonetic Continuous Speech Corpus CDROM*. 1993

[Ghahramani 1998] GHAHRAMANI, Zoubin: Learning Dynamic Bayesian Networks. In: *Adaptive Processing of Sequences and Data Structures, International Summer School on Neural Networks, Ë.R. CaianielloTutorial Lectures*. London, UK : Springer-Verlag, 1998, S. 168–197. – ISBN 3-540-64341-9

[Gilbert und Feng 2008] GILBERT, M. ; FENG, Junlan: Speech and language processing over the web. In: *Signal Processing Magazine, IEEE* 25 (2008), May, Nr. 3, S. 18–28. – ISSN 1053-5888

[Huang und Lee 1991] HUANG, X. D. ; LEE, K. F.: On speaker-independent, speaker-dependent, and speaker-adaptive speech recognition. In: *ICASSP '91: Proceedings of the Acoustics, Speech, and Signal Processing, 1991. ICASSP-91., 1991 International Conference*. Washington, DC, USA : IEEE Computer Society, 1991, S. 877–880. – ISBN 0-7803-0003-3

[Kokaram u. a. 2006] KOKARAM, A. ; REA, N. ; DAHYOT, R. ; TEKALP, M. ; BOUTHEMY, P. ; GROS, P. ; SEZAN, I.: Browsing sports video: trends in sports-related indexing and retrieval work. In: *Signal Processing Magazine, IEEE* 23 (2006), March, Nr. 2, S. 47–58. – ISSN 1053-5888

[Liu u. a. 1993] LIU, Fu-Hua ; STERN, Richard M. ; HUANG, Xuedong ; ACERO, Alejandro: Efficient cepstral normalization for robust speech recognition. In: *HLT '93: Proceedings of the workshop on Human Language Technology*. Morristown, NJ, USA : Association for Computational Linguistics, 1993, S. 69–74. – ISBN 1-55860-324-7

[Nguyen 2009] NGUYEN, P.: Techware: speech recognition software and resources on the web. In: *Signal Processing Magazine, IEEE* 26 (2009), May, Nr. 3, S. 102–105. – ISSN 1053-5888

[Ohtsuki u. a. 2006] OHTSUKI, K. ; BESSHO, K. ; MATSUO, Y. ; MATSUNAGA, S. ; HAYASHI, Y.: Automatic multimedia indexing: combining audio, speech, and visual information to index broadcast news. In: *Signal Processing Magazine, IEEE* 23 (2006), March, Nr. 2, S. 69–78. – ISSN 1053-5888

[O'Shaughnessy 2008] O'SHAUGHNESSY, D.: Invited paper: Automatic speech recognition: History, methods and challenges. 41 (2008), October, Nr. 10, S. 2965–2979

[Ostendorf 2008] OSTENDORF, M.: Speech technology and information access [In the Spotlight]. In: *Signal Processing Magazine, IEEE* 25 (2008), May, Nr. 3, S. 152–150. – ISSN 1053-5888

[Rabiner 2003] RABINER, Lawrence: COMPUTER SCIENCE: The Power of Speech. In: *Science* 301 (2003), Nr. 5639, S. 1494–1495. – URL http://www.sciencemag.org

[Ranjan u. a. 2006] RANJAN, Abhishek ; BALAKRISHNAN, Ravin ; CHIGNELL, Mark: Searching in audio: the utility of transcripts, dichotic presentation, and time-compression. In: *CHI '06: Proceedings of the SIGCHI conference on Human Factors in computing systems*. New York, NY, USA : ACM Press, 2006, S. 721–730. – ISBN 1-59593-372-7

[Ringlstetter u. a. 2007] RINGLSTETTER, Christoph ; SCHULZ, Klaus U. ; MIHOV, Stoyan: Adaptive text correction with Web-crawled domain-dependent dictionaries. In: *ACM Trans. Speech Lang. Process.* 4 (2007), Nr. 4, S. 9. – ISSN 1550-4875

[Scharenborg 2007] SCHARENBORG, Odette: Reaching over the gap: A review of efforts to link human and automatic speech recognition research. In: *Speech Commun.* 49 (2007), Nr. 5, S. 336–347. – ISSN 0167-6393

[Spina und Zue 1996] SPINA, M. S. ; ZUE, V.: Automatic Transcription of General Audio Data: Preliminary Analyses. In: *Proc. ICSLP '96* Bd. 2. Philadelphia, PA, 1996, S. 594–597

[Wang u. a. 2008] WANG, Ye-Yi ; YU, Dong ; JU, Yun-Cheng ; ACERO, A.: An introduction to voice search. In: *Signal Processing Magazine, IEEE* 25 (2008), May, Nr. 3, S. 28–38. – ISSN 1053-5888

[Witt 2008] WITT, Kristoffer: AW1: Transkription von Radiospots. (2008), December

[Zue u. a. 1989] ZUE, Victor ; GLASS, James ; PHILLIPS, Michael ; SENEFF, Stephanie: The MIT SUMMIT Speech Recognition system: a progress report. In: *HLT '89: Proceedings of the workshop on Speech and Natural Language*. Morristown, NJ, USA : Association for Computational Linguistics, 1989, S. 179–189