HOCHSCHULE DER MEDIEN

SWR

Masterarbeit

im Studiengang Audiovisuelle Medien (AM3)

an der Hochschule der Medien Stuttgart

Künstliche Intelligenz im Hörfunk

Eine Technologie der Zukunft für den öffentlich-rechtlichen Rundfunk?

Vorgelegt von:

Eva-Maria Schiemer (31744)

am 08. April 2018

zur Erlangung des akademischen Grades eines

Master of Engineering

Erstprüfer: Prof. Oliver Curdt

Zweitprüfer: Prof. Dr. Simon Wiest

Praxisbetreuer: Dipl.-Ing. Ralf Baron (SWR)

Sperrvermerk

Die vorgelegte Masterarbeit basiert auf internen, vertraulichen Daten und Informationen des Südwestrundfunks (SWR). In diese Arbeit dürfen Dritte, mit Ausnahme der Gutachter und befugten Mitgliedern des Prüfungsausschusses, ohne ausdrückliche Zustimmung des Unternehmens und des Verfassers, keine Einsicht nehmen. Eine Veröffentlichung der Masterarbeit ohne ausdrückliche Genehmigung – auch auszugsweise – ist nicht erlaubt. Ebenso dürfen keinerlei Kopien oder Abschriften – auch in digitaler Form – gefertigt werden. Ausnahmen bedürfen der schriftlichen Genehmigung des Südwestrundfunks.

Eidesstattliche Erklärung

Hiermit versichere ich, Eva-Maria Schiemer, ehrenwörtlich, dass ich die vorliegende Masterarbeit mit dem Titel: „Künstliche Intelligenz im Hörfunk - Eine Technologie der Zukunft für den öffentlich-rechtlichen Rundfunk?“ selbstständig und ohne fremde Hilfe verfasst und keine anderen als die angegebenen Hilfsmittel benutzt habe. Die Stellen der Arbeit, die dem Wortlaut oder dem Sinn nach anderen Werken entnommen wurden, sind in jedem Fall unter Angabe der Quelle kenntlich gemacht. Die Arbeit ist noch nicht veröffentlicht oder in anderer Form als Prüfungsleistung vorgelegt worden.

Ich habe die Bedeutung der ehrenwörtlichen Versicherung und die prüfungsrechtlichen Folgen (§23 Abs. 2 Master-SPO (3 Semester) der HdM) einer unrichtigen oder unvollständigen ehrenwörtlichen Versicherung zur Kenntnis genommen.

Ort, Datum	Eva-Maria Schiemer

Danksagung

An dieser Stelle möchte ich mich bei einigen Personen bedanken, die zum Gelingen dieser Arbeit beigetragen haben und mir während meines Studiums stets tatkräftig zur Seite standen:

An erster Stelle geht besonderer Dank an Ralf Baron, sowie Prof. Oliver Curdt und Prof. Dr. Simon Wiest für die Betreuung dieser Arbeit, die fachliche Unterstützung und die lehr- und erfahrungsreiche Zeit in den vergangenen Semestern an der HdM.

Des Weiteren möchte ich mich meinem Abteilungsleiter Maximilian Federhofer und allen Kollegen der Abteilung Hörfunk Sendebetrieb in Baden-Baden bedanken. Ohne den Rückhalt des Kollegenkreises ist ein solches Projekt über einen Zeitraum von sechs Monaten nicht zu stemmen. Einen großen Dank an Philip Bell und Gaby Baldowe von der SBDispo, die jeden meiner Wünsche bei der Erstellung des Dienstplanes berücksichtigt haben.

Für die Bereitschaft zu einem Experteninterview und die Bereitstellung der Studierendenlizenz möchte ich mich an dieser Stelle auch bei Frank Feulner (AX Semantics) und Aeneas Rooch (Scisys) bedanken. Sie standen mir während des gesamten Schreibprozesses mit Rat und Tat zur Seite.

Zu guter Letzt gilt der größte Dank meinen Freunden und meiner Familie, und dabei ganz besonders meinen Eltern. Sie haben mir jederzeit motivierenden Rückhalt und gleichzeitig Freiraum gegeben, um meine Ziele zu verfolgen und das Studium erfolgreich abzuschließen

Kurzfassung

Künstliche Intelligenz (KI) ist bereits jetzt in vielen Lebensbereichen nicht mehr weg zu denken. Jedoch sind die aktuellen Einsatzgebiete erst der Beginn einer Entwicklung, deren Potential in viele Disziplinen noch gar nicht absehbar ist. In wieweit der Einsatz in einem öffentlich-rechtlichen Rundfunkunternehmen sinnvoll ist, erörtert die vorliegende Arbeit. Bevor es um die Vorstellung konkreter Tools und deren Anwendungsgebiete geht, werden die technischen Hintergründe erklärt. Dem geht eine geschichtliche Einordnung der jungen Disziplin KI voraus. Im technischen Teil der Arbeit werden die Grundlagen des maschinellen Lernens, der neuronalen Netze und der Spracherkennung dargestellt. Im darauffolgenden Teil der Arbeit steht die Anwendung dieser Technologien im Mittelpunkt. An dieser Stelle geht es um die Anknüpfungspunkte verschiedener Tools in die Arbeitsabläufe eines öffentlich-rechtlichen Rundfunkunternehmens. Ein besonderes Augenmerk wurde auf die Bereiche Redaktion, Studio, Archiv und Visual Radio gelegt.

Schlüsselwörter: Künstliche Intelligenz, Rundfunk, Medien, Audio, Video, maschinelles Lernen, neuronale Netze, Spracherkennung.

Abstract

Artificial intelligence (AI) is already indispensable in many areas of life. However, current applications are only the beginning of a development whose future so far no one can give accurate information. This thesis discusses to what extent artificial intelligence is useful in public service broadcasting. Starting with a historical classification of the young discipline AI, this thesis leads on to presenting its technical background. Finally, specific tools will be presented. The technical part of the paper presents the basics of machine learning, neural networks and speech recognition. The next part is focused on the application of these technologies by connecting different tools in the workflow of a public service broadcasting company. Particular attention was paid to the areas of editorial office, studio, archive and especially visual radio.

Keywords: artificial Intelligence, broadcasting, media, audio, video, machine learning, neural network, speech recognition.

Inhaltsverzeichnis

Sperrvermerk ..... **I**

Eidesstattliche Erklärung ..... **III**

Danksagung ..... **V**

Kurzfassung ..... **VII**

Inhaltsverzeichnis ..... **IX**

Abbildungsverzeichnis ..... **A**

Tabellenverzeichnis ..... **B**

Abkürzungsverzeichnis ..... **C**

1 Einführung **1**

1.1 Der Südwestrundfunk 3

2 Künstliche Intelligenz **5**

2.1 Begriffserklärung 6

2.2 Geschichte der künstlichen Intelligenz 7

3 Maschinelles Lernen **12**

3.1 Überwachtes Lernen 12

3.1.1 Entscheidungsbäume 12

3.1.2 Instanzbasiertes Lernen 13

3.2 Unüberwachtes Lernen 14

3.2.1 Assoziative Regeln 14

3.2.2 Clusteranalyse 15

3.3 Reinforcement Learning 16

4 Neuronale Netze **18**

4.1 Natürliche Neuronale Netze 18

4.2 Künstliche Neuronale Netze 20

4.3 Deep Learning 24

5 Spracherkennung 26

5.1 Sprachverarbeitung 27

5.1.1 Schwierigkeiten in der Sprachverarbeitung 27
5.1.2 Aufbau von Spracherkennungssystemen 29
5.2 Akustische Modelle für gesprochene Sprache 30

5.2.1 Aufbau und Funktion des Sprechapparats 30
5.2.2 Das Quelle-Filter-Modell 32
5.2.3 Einheiten 33
5.3 Signalanalyse 34

5.3.1 Diskrete Fourier-Transformation 34
5.3.2 Lineare Prädiktive Codierung (LPC) 36
5.3.3 Cepstrum 38
5.4 Akustische Bewertung 39

5.4.1 Markov-Kette 40
5.4.2 Probleme der HMMs 41
5.5 Dekodierung 41

5.5.1 Fehlerarten 42

6 Stand der Technik 44

6.1 IBM Watson 44

6.1.1 Jeopardy! 44
6.1.2 Trailer für Horrorfilm „Morgan“ 45
6.1.3 Cognitive Highlights 46
6.2 Smartphone-Technologie 46

6.2.1 Huawei Mate 10 Pro 46
6.2.2 iPhone X 47
6.3 Adobe VoCo 47

6.4 Google Clips 49

6.5 Audio Mining 50

6.6 Wibbitz 51

7 Tools im Anwendungstest 52

7.1 Roboterjournalismus 52

7.1.1 NLG-Cloud 53
7.1.2 AX Semantics University 59
7.2 Automatisierter Videoschnitt 60
7.2.1 Mevo-Kamera 60
7.3 Automatisierter Audioschnitt 61
7.3.1 Scisys 62
7.4 Cloud-Computing 64
7.4.1 IBM: Bluemix 64
7.4.2 Microsoft: Azure 71
7.4.3 Auswertung 79
8 Anwendungsbereiche im Hörfunk am Beispiel des SWR 82
8.1 Status Quo im SWR 82
8.2 Archiv 82
8.3 Redaktion 83
8.4 Sendestudio 83
8.5 Visual Radio 85
9 Fazit 91
Literaturverzeichnis ..... E
Quellenverzeichnis I
Anhang i
1. IPA-Zeichen i
2. Technische Daten Smartphones ii
a. Huawei Mate 10 Pro (Auszug) ii
b. iPhone X (Auszug) iii
3. Interview Frank Feulner (AX Semantics) iv
4. Technische Daten: Mevo-Kamera (Auszug) xiii
5. Interview Aeaneas Rooch (Scisys) xiv
6. IBM Bluemix: Cloud-Services (Auszug) xxi

7. Ergebnisse IBM Speech-to-Text xxii

8. Bilder für IBM Watson Visual Recognition xxiv

9. Microsoft Azure: AI und Cognitive Services xxvi

10. Ergebnisse Azure BING-Spracheingabe xxvii

Abbildungsverzeichnis

Abbildung 2-1: Feed-Forward-Architektur 8
Abbildung 3-1: Eine Punktmenge in einem Voronoi-Diagramm 14
Abbildung 3-2: Ergebnis einer Clusteranalyse mit Normalverteilung. 16
Abbildung 3-3: Beziehung zwischen Agent und Umwelt 17
Abbildung 4-1: Modell eines biologischen Neurons im Gehirn 19
Abbildung 4-2: Schema eines künstlichen Neurons. 20
Abbildung 4-3: Rezeptorendarstellung eines Neurons 21
Abbildung 4-4: Stufenfunktion und Sigmoide Funktion im Vergleich. 23
Abbildung 4-5: Rezeptorschichtdarstellung 23
Abbildung 4-6: Gesichtserkennung via Deep Learning 25
Abbildung 5-1: Prinzipieller Aufbau von Spracherkennern. 29
Abbildung 5-2: Zuordnung der linearen Frequenzskalen zur Bark- und Mel-Skala 36
Abbildung 5-3: Approximation eines Sprachspektrums durch Prädiktoren 38
Abbildung 5-4: HMM 1. Ordnung mit N Zuständen 40
Abbildung 5-5: Anfang des Modells für das Wort "acht" mit ε-Zuständen 42
Abbildung 6-1: Funktionsweise Adobe VoCo 48
Abbildung 6-2: Google Clips-Kamera 49
Abbildung 7-1: Bestandteile der Textgenerierung mit der NLG-Cloud. 54
Abbildung 7-2: Datenanalyse mit Fehlern (markierter Bereich). 55
Abbildung 7-3: Datenanalyse mit Korrektur (markierter Bereich). 55
Abbildung 7-4: Von den Daten zum Text – die einzelnen Schritte 56
Abbildung 7-5: Funktionsweise der Property – true. 57
Abbildung 7-6: Funktionsweise der Property – false 57
Abbildung 7-7: Mevo-Kamera 61
Abbildung 7-8: Vergleich Spracherkennung IBM <-> Azure. 80
Abbildung 7-9: Vergleich Bilderkennung IBM <-> Azure. 81
Abbildung 8-1: Signalweg 84
Abbildung 8-2: Kameraeinstellungen des Visual Radio mit Einblendungen 85
Abbildung 8-3: Schematische Darstellung der Weiterentwicklung des Visual Radio. 87
Abbildung 8-4: Amazon Echo Show 89

Tabellenverzeichnis

Tabelle 1: Vergleich der Rechenressourcen.....6

Tabelle 2: Definition der Variablen – nach Norbert Hoffmann.....21

Tabelle 3: Elementare Bausteine in der Wortzusammensetzung.....27

Tabelle 4: Unterscheidungskriterien für Spracherkennung.....29

Tabelle 5: Übersicht der an der Sprachproduktion beteiligten Organe und deren Funktion.....30

Tabelle 6: Übersicht der sprachlichen Einheiten.....34

Tabelle 7: Vgl. Datenanalyse (redaktionell/technisch).....54

Tabelle 8: Gegenüberstellung der Anforderungen an einen Text und an viele Texte.....58

Tabelle 9: Bedingungen bei Properties.....58

Tabelle 10: Übersicht der Unternehmen, deren KI und Cloud.....64

Tabelle 11: Testdurchläufe bei Speech-to-Text.....66

Tabelle 12: Ergebnisübersicht IBM Speech-to-Text.....68

Tabelle 13: Ergebnisübersicht Azure BING-Spracheingabe.....74

Abkürzungsverzeichnis

ARD	Arbeitsgemeinschaft der öffentlich-rechtlichen Rundfunkanstalten in der Bundesrepublik Deutschland
ATML	Automated Text Markup Language
AWS	Amazon Web Services
DFKI	Deutsches Forschungszentrum für Künstliche Intelligenz
FESAD	Fernseharchivdatenbank
HdM	Hochschule der Medien
HF	Hörfunk
HFDB	Hörfunkdatenbank
IoT	Internet of Things (dt. Internet der Dinge)
KiKa	Kinderkanal
KDD	Knowledge Discovery in Databases
KI	Künstliche Intelligenz
KNN	künstliche neuronale Netze
Ladar	Laser Detecting and Ranging
Lidar	Light Detecting and Ranging
LMU	Ludwig-Maximilians-Universität
NLG	Natural Language Generation
NDB	Normdatenbank
SWR	Südwestrundfunk
WA	Wortakkuratheit
WK	Wortkorrektheit

1 Einführung

Dass es schon bald selbstfahrende Autos geben wird, steht außer Frage. Schon seit einer geraumen Zeit arbeiten große Automobilunternehmen mit Zulieferern wie Bosch, aber auch Google und Facebook mit Hochdruck an automatisierten Fahr- und Fahrassistenzsystemen. Künftig werden unsere Straßen von Autos befahren, die mittels Kameras und Sensoren (Lidar[1] und Ladar[2]) ein 3D-Bild der Autoumgebung erzeugen und so Straßen und andere Verkehrsteilnehmer erkennen und entsprechend reagieren können.

Der Bereich des automatisierten Fahrens ist allerdings bei weitem nicht das einzige Einsatzgebiet der künstlichen Intelligenz. In der Medizin werden Systeme entwickelt, die eigenständig die Befragung des Patienten übernehmen und die Diagnose stellen können. Durch die Bilderkennung wird die Auswertung von medizinischem Bildmaterial (Röntgenbilder, Ultraschall, …) von Maschinen übernommen. Dabei werden Ärzte in Zukunft hauptsächlich die Aufgabe haben, die Diagnosen der künstlichen Intelligenz zu überprüfen und die Behandlungsvorschläge freizugeben. Auch Operationsroboter werden den Weg in die Krankenhäuser finden.

Aber auch gelähmte Menschen können sich in den nächsten Jahrzehnten Hoffnungen machen, ihre Bewegungsfreiheit zurückzuerlangen. In der BrainGate2-Studie um Dr. Bolu Ajiboye wurde bei einem Tetraplegiker[3] über eine Schnittstelle das Gehirn mit den Muskeln an den Armen verbunden. Zur Überbrückung und Übermittlung dieser Signale dienten Kabelwege. Dazu wurden dem Mann während einer Operation Sensoren in die Region des Gehirns gepflanzt, die für die Bewegung zuständig ist. Diese Sensoren wurden mit einem Computer verbunden. In einer viermonatigen Phase lernte dieser Computer, welche Hirnsignale welche Bewegung auslösten. Dazu musste sich der Patient die entsprechende Bewegung vorstellen und in seinen Gedanken ausführen. Dann wurden dem Patienten 36 muskelstimulierende Elektroden in den Arm eingepflanzt. Um die über die Jahre verlorengegangene Kraft der Armmuskulatur wieder aufzubauen, wurden die betroffenen Muskeln zunächst durch die Elektroden angeregt. Danach wurden die Muskeln mit der Hirn-

[1] Lidar: engl. Light Detecting and Ranging.
[2] Ladar: engl. Laser Detecting and Ranging.
[3] Tetraplegie: Form der Querschnittslähmung, bei der alle vier Gliedmaßen von der Lähmung betroffen sind.

Computer-Schnittstelle verbunden. Dadurch konnten nun die Hirnsignale in Muskelkontraktion übersetzt werden.[4]

Ähnliche Forschungen unternimmt Facebook derzeit. Ziel der Forschungsarbeit ist es, Textnachrichten mittels Gedanken zu verschicken. Dazu muss ein Computer die Übersetzung von Gehirnsignalen in Buchstaben lernen, bevor diese auf das Smartphone übertragen werden. Ist die Forschungsarbeit erfolgreich, könnten Nachrichten verschickt werden, ohne dass sie eingetippt werden müssen. Dieser Gedankenaustausch soll jedoch nicht über eingepflanzte Elektroden, sondern über einen Sensor auf dem Kopf möglich sein.[5]

Allerdings eröffnen diese Möglichkeiten einen großen Katalog an Fragen, die im Vorhinein beantwortet werden müssen. Dies betrifft allen voran ethische und moralische Fragestellungen. Die Diskussion dieses Katalogs ist jedoch nicht Gegenstand dieser Arbeit. Stattdessen soll es um die unterschiedlichen Anwendungsmöglichkeiten der künstlichen Intelligenz in einem modernen Medienunternehmen gehen.

Im Bereich des Rundfunks werden spezielle Anforderungen an eine solche Software gestellt, die nicht immer einfach umzusetzen sind. Zum einen ist wichtig, dass die Software den Umgang mit der deutschen Sprache beherrscht. Da es sich beim SWR um eine öffentlich-rechtliche Sendeanstalt in Deutschland handelt, ist beispielsweise eine Bereitstellung von deutschen Texten oder die Spracherkennung deutscher Sprache Voraussetzung für einen gelingenden Einsatz. Bis jetzt scheitert die Implementierung der deutschen Sprache jedoch oft an der komplexen Grammatik. Eine Sprach- oder Texterkennung in Englisch arbeitet schon jetzt zuverlässig. Zum anderen muss die Software ein gewisses Maß an Flexibilität zulassen. Der SWR hat eine ganze Reihe verschiedener Zielgruppen, die alle gleichermaßen bedient werden müssen. Zudem darf die Ausfallsicherheit nicht außer Acht gelassen werden.

Mithilfe der künstlichen Intelligenz wurden von diversen Unternehmen bereits jetzt eine Vielzahl von Tools und Anwendungen konzeptioniert und umgesetzt, die speziell im Rundfunk für eine

---

[4] Vgl.: A Bolu Ajiboye, Francis R. Willett, Daniel R. Young, Robert F. Kirsch, u.a.: Restoration of reaching and grasping movements through brain-controlled muscle stimulation in a person with tetraplegia: a proof-of-concept demonstration. In: *The Lancet* 389 (10081), London 2017, S. 1821–1830.

[5] Vgl.: „Facebook will Gedanken in Text umwandeln“, Mainz 2017. http://www.3sat.de/page/?source=/kulturzeit/news/192284/index.html (Stand: 13.11.2017)

große Entlastung des Arbeitspensums sorgen. So konnte die Geschwindigkeit der Arbeitsabläufe immens erhöht und der Aktualität somit ein ganz neuer Stellenwert gegeben werden.

In welchem Rahmen diese Anwendungen die Arbeitsabläufe in einem öffentlich-rechtlichen Rundfunksender optimieren können, diskutiert die vorliegende Masterthesis. Zu Beginn steht ein Einblick in die Geschichte der künstlichen Intelligenz. Dem folgen die technischen Grundlagen des maschinellen Lernens, der neuronalen Netze und der Spracherkennung. Während Kapitel 6 erste Rechercheergebnisse näher beleuchtet, werden besonders interessante Tools in Kapitel 7 vorgestellt und getestet. Die Ergebnisse jedes Tools werden auf eine Anwendung in einem Rundfunkunternehmen, wie dem Südwestrundfunk, diskutiert. Wichtige Themenfelder sind dabei die Effizienz und Benutzerfreundlichkeit der Tools.

Das abschließende Fazit gibt eine Übersicht der Ergebnisse und stellt diese in einen größeren Kontext. Ein Blick in die Zukunft der künstlichen Intelligenz rundet die vorliegende Arbeit ab.

1.1 Der Südwestrundfunk

Der Südwestrundfunk, im Folgenden „SWR“ genannt, ist die zweitgrößte der insgesamt neun Landesrundfunkanstalten der „Arbeitsgemeinschaft der öffentlich-rechtlichen Rundfunkanstalten der Bundesrepublik Deutschland“ (ARD).

Das Sendegebiet des SWR erstreckt sich über die Bundesländer Baden-Württemberg und Rheinland-Pfalz. Der SWR erreicht mit seinem länderübergreifenden Fernsehprogramm und den acht Hörfunkprogrammen (SWR 1 BW, SWR 1 RP, SWR 2, SWR 3, SWR 4 BW, SWR 4 RP, DASDING, SWR Aktuell) rund 15 Millionen Menschen.

Deutschlandweit können die Menschen den SWR über die Sendungen der ARD-Gemeinschaftsprogramme empfangen: Das Erste, 3sat, ARTE, Phoenix, KiKA, funk, One, ARD-Alpha und Tagesschau24. Die etwa 3800 festangestellten Mitarbeiter sind auf Funkhäuser und (Regional-)Studios im gesamten Sendegebiet verteilt.

Mit seinen Regionalstudios ist der SWR einerseits nah an den Menschen, führt andererseits ein Informationsnetzwerk weltweit verteilter Korrespondenten. Mit diesen beiden Komponenten will der SWR dem Auftrag eines öffentlich-rechtlichen Rundfunksenders gerecht zu werden. Damit hat

er sich seit seiner Gründung (nach der Fusion von SDR und SWF im Jahr 1998) zu einem modernen Medienunternehmen mit regionalem Profil entwickelt.[6]

Das junge Programm des SWR, der Sender DASDING, zeigt besonderes Interesse an innovativen Themen und zukunftsfähigen Entwicklungen im Bereich des Hörfunks. On Air werden Themen rund um Livestyle und Mode, Internet und Multimedia, sowie Nachrichten und Job/Karriere behandelt. Aber auch der Bereich aktueller technischer Entwicklungen nimmt einen großen Stellenwert ein. Dies spielt sowohl on Air, als auch im Hintergrund eine große Rolle.

Die Einführung des Visual Radio (nähere Informationen in Kapitel 8.5) im Jahr 2014 war ein großer Schritt in diese Richtung. Dabei wird das Hörfunkprogramm in Echtzeit über einen Video-Livestream ins Internet übertragen. Dadurch wurde für die Zielgruppe der 14- bis 29-jährigen eine neue Art der Kommunikation und Einbindung angeboten. Um dieses Angebot weiter auszubauen, möchte der Teamleiter der DASDING-Techniker Dipl.- Ing. Ralf Baron in Zukunft auch auf Technologien im Bereich der künstlichen Intelligenz zurückgreifen.

[6] Vgl. Wolfgang Utz: SWR – Zahlen Daten Fakten. Stuttgart 2017.

2 Künstliche Intelligenz

Elaine Rich definierte im Jahr 1983 die künstliche Intelligenz als

> *„the study of how to make computers do things at which, at the moment, people are better".*[7]

Für den Aufbau einer künstlichen Intelligenz hat sich das menschliche Gehirn als geeignetes Vorbild erwiesen. Zu dieser Zeit waren uns die Computer besonders im Bereich der Lernfähigkeit weit unterlegen. Um die Differenz auszugleichen, mussten verschiedene Verfahren des maschinellen Lernens weiterentwickelt werden – beruhend auf unterschiedlichen wissenschaftlichen Disziplinen.

Denn die künstliche Intelligenz ist ein weites Feld, das viele Fachrichtungen tangiert und miteinander verbindet. Die Sprachwissenschaft spielt bei der Interpretation von natürlicher Sprache durch Maschinen eine wesentliche Rolle. Die Psychologie beschäftigt sich bereits seit Jahrhunderten mit dem Intelligenzbegriff (Kapitel 2.1) und den Fragen in Bezug auf Denken und Handeln bei Mensch und Tier. Die Philosophie spielt im Bereich des logischen Schließens eine große Rolle, ebenso bei der Frage nach dem Ursprung des Wissens und dessen Übersetzung in eine Aktion.

In der Informatik kommen Fragestellungen der Philosophie und mathematische Regeln zusammen und werden in Algorithmen verarbeitet. Die Neurowissenschaften sind bei der Bildung geeigneter Neuronen-Modelle beteiligt: Sie liefern Daten über die einzelnen Bereiche und deren Verbindungen im Gehirn. Obwohl Forscher mittlerweile in der Lage sind, einzelne Neuronen elektrisch, chemisch oder optisch zu stimulieren, sind die Neurowissenschaften immer noch weit davon entfernt, kognitive Prozesse nachzuvollziehen.[8]

In Tabelle 1 sind die Ressourcen eines IBM BlueGene-Supercomputers, einem typischen Personalcomputer und eines menschlichen Gehirns gegenübergestellt. Da sich die Performance des Supercomputers etwa alle fünf Jahre verzehnfachten, konnten sie sich denen eines Gehirns immer weiter annähern.[9]

---

[7] Elaine Rich, Kevin Knight, Shivashankar B. Nair: Artificial Intelligence. Neu-Delhi 2009, S. 3.
[8] Vgl. Stuart Russel, Peter Norvig: Künstliche Intelligenz. Ein moderner Ansatz. München 2012, S. 27ff.
[9] Vgl. Russell, Norvig: Künstliche Intelligenz, S. 34.

	Supercomputer	Personalcomputer	Menschliches Gehirn
Recheneinheiten	10^4 CPUs, 10^{12} Transistoren	4 CPUs, 10^9 Transistoren	10^{11} Neuronen
Speichereinheiten	10^{14} Bit RAM	10^{11} Bit RAM	10^{11} Neuronen
	10^{15} Bit Festplatte	10^{13} Bit Festplatte	10^{14} Synapsen
Zykluszeit	10^{-9} s	10^{-9} s	10^{-3} s
Operationen/s	10^{15}	10^{10}	10^{17}
Speicheraktualisierungen/s	10^{14}	10^{10}	10^{14}

Tabelle 1: Vergleich der Rechenressourcen.
(Quelle: Russel, S., Norvig, P.: „Künstliche Intelligenz. Ein moderner Ansatz“ (S. 34)).

Allerdings sind Computer dem menschlichen Gehirn nicht in allen Bereichen überlegen. Deutlich ist der Unterschied beim Erkennen von Zusammenhängen: Hier sind die menschlichen Fähigkeiten denen eines Computers weit unterlegen. Geht es allerdings um Alltagsprobleme, kann der Mensch diese schneller lösen. Um alle Fähigkeiten des Menschen in einer Maschine nachzubauen bzw. zu programmieren reichen die bisherigen Erkenntnisse nicht aus. Anders sieht es jedoch bei sogenannten „Expertensystemen“ aus, welche Aufgaben eines speziellen Gebiets (z.B. Schach) übernehmen können, und innerhalb dieses Gebiets zu Höchstleistungen fähig sind. Sie sind bereits weit verbreitet.

2.1 Begriffserklärung

Der Begriff der künstlichen Intelligenz, engl. „Artificial Intelligence“, lässt sich nur sehr schwer fassen und erklären. Während „künstlich“ hauptsächlich als das Gegenteil von natürlich angesehen wird, gestaltet sich das Problem beim Intelligenzbegriff als wesentlich schwieriger.

Charles Spearman[10] veröffentlichte 1904 seine *Zwei-Faktoren-Theorie*[11], bei der nach einer Faktorenanalyse mittels eines Tests ein g-Faktor („general intelligence“) und mehrere s-Faktoren („specific intelligence“) bestimmt werden. Während der g-Faktor über alle Leistungsbereiche berechnet wird, repräsentieren die s-Faktoren die Intelligenz für einzelne, unabhängig voneinander zu bewältigende Aufgaben.

[10] Charles Spearman (1863-1945): brit. Psychologe.
[11] Vgl. Charles Spearman: ’General intelligence’ objectively determined and measured. In: American Journal of Psychology 15-2, Illinois 1904, S. 201ff.

Obwohl die Zwei-Faktoren-Theorie von Spearman von führenden Wissenschaftlern mittlerweile als nicht mehr zeitgemäß angesehen wird, hat sich dennoch die Idee eines allgemeinen Intelligenzfaktors durchgesetzt.

Demgegenüber steht das *Primärfaktorenmodell* von Luis Leon Thurstone[12]: Er sieht die Intelligenz als eine Zusammensetzung von sieben einzelnen Fähigkeiten. Diese Primärfaktoren sind: räumliches Vorstellungsvermögen, Rechenfähigkeit, Sprachverständnis, Wortflüssigkeit, Gedächtnis, Auffassungsgeschwindigkeit und logisches Denken.[13]

Ausgehend von Spearmans Modell entwickelte Raymond Bernard Cattell[14] das *Zwei-Faktoren-Modell.* Er unterschied zwischen fluiden und kristallinen Intelligenzfaktoren. Dabei werden alle angeborenen bzw. vererbten Fähigkeiten, die nicht von der Umwelt beeinflusst werden, als *fluid* bezeichnet. *Kristalline* Faktoren sind demgegenüber alle Fähigkeiten, die im Laufe des Lebens erlernt und durch die Umwelt beeinflusst werden.

Um eine allgemeinverbindliche Definition für den Begriff „Intelligenz" zu finden, müssen also mehrere Indikatoren berücksichtigt werden. Dennoch sind zweifellos kognitive Fähigkeiten, sowie Abstraktionsvermögen, Erkenntnisfähigkeit einschließlich der Fähigkeit, diese Erkenntnisse zu kommunizieren, wichtige Merkmale von Intelligenz.[15]

Bevor nun im weiteren Verlauf der Arbeit die technischen Grundlagen vorgestellt werden, ist es wichtig einen Blick in die Anfängen dieser Wissenschaft zu werfen.

2.2 Geschichte der künstlichen Intelligenz

Bereits 1943 haben sich Warren McCulloch und Walter Pitts mit dem Thema Künstliche Intelligenz beschäftigt[16]. Sie schufen ein Modell mit künstlichen Neuronen, bei dem jedes Neuron die Eigenschaft „an" oder „aus" annehmen konnte. Das Neuron konnte in den Status „an" wechseln, sobald genügend benachbarte Neuronen „an" waren und so das beobachtete Neuron stimuliert werden konnte. McCulloch und Pitts konnten damit nachweisen, dass jede berechenbare Funktion von einem Netz aus verbundenen Neuronen berechnet werden kann, und dass alle logischen Ver-

[12] Louis Leon Thurstone (1887-1955): amerik. Ingenieur und Psychologe.
[13] Vgl. George A. Lundberg: Primary Mental Abilities. L. L. Thurstone. In: American Journal of Sociology 44-2. Chicago 1938, S. 310f.
[14] Raymond Bernard Cattell (1905-1998): brit.-amerik. Persönlichkeitspsychologe.
[15] Vgl. Peter Zöller-Greer: Künstliche Intelligenz. Grundlagen und Anwendungen. Wächtersbach 2010, S. 5.
[16] Vgl. Warren McCulloch, William Pitts: A logical calculus of the ideas immanent in nervous activity. In: Bulletin of Mathematical Biophysics, Bd. 5 (1943), S. 115–133.

knüpfungen (UND, ODER, NICHT, usw.) durch einfache Netzstrukturen implementiert werden können. Dies alles führte die beiden Forscher zu der Behauptung, dass geeignet definierte Netze lernfähig seien. Ihre Arbeit gilt damit als die erste Arbeit, die allgemein als KI anerkannt wurde.

Anfang der 1950er Jahre bauten Marvin Minsky und Dean Edmons an der Harvard University den ersten neuronalen Netzcomputer. Dieser bestand aus 3000 Vakuumröhren, sowie einem alten Autopilotmechanismus aus einem B-24-Bomber. Ziel der beiden Studenten war die Simulation eines Netzes aus 40 Neuronen. Der Name des Netzcomputers war SNARC.[17]

Ebenso befasste sich Alan Turing mit der künstlichen Intelligenz. Er formulierte bereits 1950 den bis heute gültigen „Turing-Test". Mithilfe dieses Tests kann festgestellt werden, ob ein Computer einen ähnlichen Verstand besitzt, wie ein Mensch. Der Test gilt als bestanden, sobald ein Mensch auf seine schriftlich formulierten Fragen hin nicht unterscheiden kann, ob die Antworten von einem Menschen oder einem Computer stammen.

Frank Rosenblatt erfand 1962 das Perzeptron, ein lernfähiges Netzwerk und somit eine Weiterentwicklung der Ideen von Pitts und McCulloch aus den frühen 1940er Jahren. Die Maschine arbeitete nach der *Feed-Forward-Architektur*, d.h. Input-Neuronen aktivieren Output-Neuronen und nicht umgekehrt.

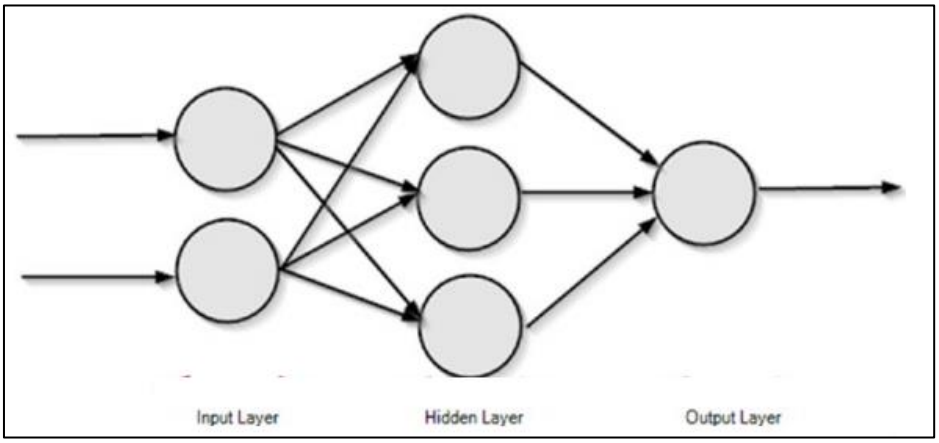


Abbildung 2-1: Feed-Forward-Architektur
(Quelle: https://dzone.com/articles/feed-forward-neural-network-with-mxnetr (Stand: 22.03.2018)).

Eine weitere wichtige Erfindung Rosenblatts waren die Verbindungsstärken (sogenannte „Gewichte"), mit denen Neuronen andere Neuronen aktivieren. Damit kann ein Perzeptron überwacht lernen.

[17] Vgl. Russell, Norvig: Künstliche Intelligenz, S. 40.

Der Optimismus dieser Zeit war in den vielversprechenden Ergebnissen früherer KI-Systeme begründet. Sollten diese Programme allerdings um Funktionalitäten erweitert werden, scheiterten sie – beispielsweise bei der „Übersetzung" vom Englischen ins Russische.[18] Der Idee zufolge sollte eine einfache syntaktische Übertragung, d.h. der schlichte Austausch von Wörtern aus einem elektronischen Wörterbuch, möglich sein. So wurde aus „Der Geist ist willig, aber das Fleisch ist schwach" im Russischen „Der Wodka ist gut, aber das Fleisch ist verfault". Der Grund liegt in dem Kontextwissen, das erforderlich ist, um Mehrdeutigkeiten zu vermeiden. Nur so kann der Inhalt der Sätze korrekt übersetzt werden.

Nach mehreren Erfahrungen dieser Art wurden alle finanziellen Unterstützungen der US-Regierung für akademische Übersetzungsprojekte eingestellt. Diesem Vorstoß folgte bald die britische Regierung: Sie traf auf Grundlage eines Berichtes von Sir James Lighthill[19] die Entscheidung, die KI-Forschung in allen bis auf drei Universitäten des Landes einzustellen. Grund dafür war die Komplexität der KI-Probleme. Funktionierten sie im Kleinen, war noch lange nicht gesagt, dass die Skalierung auf größere Probleme immer noch funktionierte. Das schlichte Ausprobieren der verschiedenen Kombinationen funktionierte nur anfangs und wurde infolge der komplexer werdenden Probleme ineffektiv.

Als Japan den Zehnjahresplan für den Aufbau intelligenter Computer unter PROLOG mit dem Titel „5. Generation" vorstellte, rief dieser Vorstoß eine Reihe von Reaktionen in unterschiedlichen Staaten hervor:

- **USA**: Gründung der „Microelectronics and Computer Technology Corporation" (MCC) im Jahr 1982 mit Hauptsitzt in Austin (Texas).
- **Großbritannien**: Unterstützung der Universitäten wird wieder aufgenommen.
- **Deutschland**: Gründung des „Deutschen Forschungszentrum für Künstliche Intelligenz" (DFKI) in Kaiserlautern 1988. Das weltweit größte Forschungszentrum mit öffentlichen und privaten Partnern forscht seit dem an den Standorten Kaiserslautern, Saarbrücken, Bremen und Berlin im Bereich innovativer Softwaretechnologien.[20]

[18] Vgl. Russell, Norvig: Künstliche Intelligenz, S. 44f.
[19] Vgl. James Lighthill: „Artifical Intelligence: A general survey". In: „Artificial Intelligence: A paper symposium", Science Research Council 1973.
[20] Weitere Informationen zum DFKI unter: https://www.dfki.de (Stand: 10.03.2018).

Im Jahr 1982 wurde „R1“ das erste erfolgreiche Expertensystem. Das Unternehmen Digital Equipment Corporation (DEC) konnte im Jahre 1986 so rund 40 Millionen Dollar einsparen. Bereits zwei Jahre später setzte das Unternehmen 40 Expertensysteme in unterschiedlichen Bereichen ein. Ende der 1980er Jahre hatte nahezu jedes größere Unternehmen eine eigene KI-Abteilung oder arbeitete an der Weiterentwicklung der Expertensysteme.

Infolge all dieser positiven Entwicklungen stieg der Umsatz der KI-Industrie von einigen Millionen im Jahre 1988 bis zu mehreren Milliarden US-Dollar an. Verschiedenste Expertensysteme, Erkennungssysteme und Roboter wurden von unzähligen Unternehmen gebaut.

Jedoch folgte bald die Periode des „KI-Winters“: Da viele Versprechen der Unternehmen nicht eingehalten werden konnten, gingen diese Konkurs und verschwanden von der Oberfläche.

Durch diese Vorkommnisse haben sich verschiedene Forschergruppen wieder auf die Anfänge der KI-Forschung zurückbesinnt. Denn bereits im Jahre 1969 hatten Bryson und Ho zum ersten Mal den *Backpropagation*-Lernalgorithmus vorgestellt. Diese Methode wurde im selben Jahr vom Finnen Seppo Linnainmaa beschrieben. Sie berechnet die nötigen Anpassungen der Gewichte in einem künstlichen neuronalen Netz, wenn dieses einen Fehler macht. Die Gewichte werden automatisch aufgrund der Ausgabe und des Vergleichs mit dem Ergebnis berechnet.

Während in den zurückliegenden Jahrzehnten eine Vielzahl neuer Modelle und Ansätze ausprobiert wurde, war es ab diesem Zeitpunkt üblich, auf existierenden Ansätzen aufzubauen und diese weiterzuentwickeln. Dazu wurden wissenschaftliche Methoden auf empirische Experimente angewandt, um in Forscherkreisen ebenso wie gesellschaftlich akzeptiert zu werden. Die im Folgenden aufgezeigten Beispiele verdeutlichen diese Entwicklung:

1. Beispiel: Spracherkennung. In den 1970er-Jahren wurden unzählige Modelle entwickelt. Diese waren oft instabil und nur für wenige, spezielle Anwendung programmiert. Gegen Ende des Jahrzehnts hat sich das Hidden-Markov-Modell (HMM) als besonders geeignet erwiesen.

2. Beispiel: Maschinelle Übersetzung. Bereits in den 1950er Jahren wurde ein Konzept (SYSTRAN von Dr. Peter Toma) entwickelt, das auf Wortsequenzen beruht. Dieses geriet danach jedoch für längere Zeit in Vergessenheit. Erst knapp 20 Jahre später wurde dieses Konzept wieder aufgefunden und führt jetzt das Gebiet an.

3. Beispiel: Data Mining. Eine ähnliche Vergangenheit hat auch das Data Mining. Die Grundlage bildeten die Entdeckungen von Warren McCulloch und Walter Pitts (1943). Da das Modell der künstlichen Neuronalen Netze (KNN) danach in Vergessenheit geriet, wurden diese Netze erst in den 1980er Jahren durch verbesserte Methodik weiterentwickelt.[21]

Seit Beginn der KI-Forschung lag der Fokus immer auf der Entdeckung der besten Algorithmen. Nach 60 Jahren KI-Geschichte hat sich nun allerdings herausgestellt, dass es oft sinnvoller ist, den Fokus auf die Daten zu legen, und weniger differenziert den Algorithmus zu betrachten. Dieses Prinzip ist aufgrund der stets wachsenden Datenmengen effizienter. Nach Banko und Brill (2001) gibt ein mittelmäßiger Algorithmus mit 100 Millionen Daten ein besseres Ergebnis aus, als der beste Algorithmus mit nur einem Anteil von 1/100 der Daten.[22]

Durch diese und ähnliche Arbeiten[23] wird deutlich, dass geeignete Lernmethoden (sofern genügend Daten zur Verfügung stehen) im Bereich der Wissensverarbeitung effektiver arbeiten als manuell kodierte Anwendungen.

[21] Vgl. Russell, Norvig: Künstliche Intelligenz, S. 39-52.
[22] Vgl. Michele Banko, Eric Brill: Scaling to Very Very Large Corpora for Natural Language Disambiguation. Redmond 2001.
[23] Für weiterführende Informationen: James Hays, Alexei A. Efros: Scene Completion Using Millions of Photographs. New York 2007, S. 87ff.

3 Maschinelles Lernen

Das Gebiet des maschinellen Lernens ist ein Teilgebiet der KI und eng verbunden mit „Data Mining“ und „Knowledge Discovery in Databases“ (KDD). Dabei geht es vorwiegend um die „computergestützte Modellierung und Realisierung von Lernphänomen, der Extraktion von Wissen aus großen Datenbeständen und der Realisierung adaptiver Systeme.“[24] Diese Verfahren sorgen bei Suchmaschinen für eine gute Sortierung, sie lernen die Präferenzen ihrer User kennen und arbeiten u.a. als Spamfilter in den E-Mailkonten.

Beim maschinellen Lernen wird zwischen überwachtem und unüberwachtem Lernen, sowie dem Reinforcement Learning unterschieden:

3.1 Überwachtes Lernen

Beim überwachten Lernen wird ein Agent auf Basis einer festen Trainingsmenge an Musterpaaren trainiert. Die Gewichte werden so optimiert, dass die Differenz zwischen Soll- und Istausgabe minimiert wird. Dies kann mittels verschiedener Methoden geschehen. Im Folgenden werden exemplarisch einfach umsetzbare Entscheidungsbäume und der „Nearest-Neighbour“-Algorithmus vorgestellt.

3.1.1 Entscheidungsbäume

Entscheidungsbaumverfahren sind seit vielen Jahrzehnten äußert populär. Sie zählen aufgrund ihrer einfachen Bedienung, leicht für Menschen nachvollziehbarer Lernergebnisse und effizienter Laufzeiten zu den benutzerfreundlichsten Verfahren.

Zum Klassifizieren wird ein Entscheidungsbaum immer von oben nach unten (bzw. von links nach rechts) durchlaufen. Der Startpunkt liegt dabei immer an der Wurzel des Baumes. An dessen Knoten wird das Attribut geprüft. Die Kanten repräsentieren den Attributwert. Nach diesem Muster wird der Entscheidungsbaum Knoten für Knoten (bzw. Attribut für Attribut) durchgelaufen, bis an deren jeweiligen Enden ein Blatt ohne weitere ausgehende Kanten steht. Dort kann der Klassenwert abgelesen werden.

Problematisch ist allerdings die hohe Rechenzeit, die mit zunehmenden Attributen einhergeht. Da diese innerhalb des Entscheidungsbaums in jedem Pfad in beliebiger Reihenfolge vorkommen kön-

[24] Stefan Wrobel, Thorsten Joachims, Katharina Morik: Maschinelles Lernen und Data Mining. In: Görz, G., Schneeberger, J., Schmid, U.: Handbuch der Künstlichen Intelligenz. Oldenburg 2014, S. 405.

nen, wächst die Anzahl der möglichen Entscheidungsbäume exponentiell mit der Anzahl der Attribute. Deshalb könnten zwar mittels eines Algorithmus alle möglichen Entscheidungsbäume erzeugt werden und derjenige mit der minimalen Fehlerzahl ausgewählt werden. Dieses Verfahren ist allerdings nicht praktikabel.

Stattdessen hat sich das „top-down induction of decision trees"-Prinzip bewährt. Dabei wird vorher entschieden, welches Attribut für den Wurzelknoten den höchsten Informationsgewinn und damit die größte Verbesserung des Klassifikationsquotienten liefert. Dieser Prozess wird für jeden Knoten wiederholt. Ziel ist es, einen möglichst kleinen und effizienten Entscheidungsbaum zu produzieren. Durch Techniken wie *Pruning*[25], kann der Entscheidungsbaum nachträglich verkürzt werden.

3.1.2 Instanzbasiertes Lernen

Unter „instanzbasiertem Lernen" versteht man die Verfahren des „*Nearest Neighbour*". Grundlage bildet der Vergleich zwischen bereits klassifizierten und neuen Objekten. Unbekannte Objekt werden dabei nach Ähnlichkeit klassifiziert und in die Klasse eingeordnet, deren Repräsentanten ihm am ähnlichsten ist.

Anders als bei Entscheidungsbäumen wird hier keine Hypothese gebildet, sondern die Beispiele erst abgespeichert. Im nächsten Schritt werden bei der Überprüfung eines neuen Objekts die Daten miteinander verglichen. Dadurch liegt die Performance oft über der von Entscheidungsbäumen.

Der Vergleich basiert also auf zwei Objekten. Diese sind sich umso ähnlicher, je geringer der Abstand ihrer Merkmale ist. Dieser Abstand kann im einfachsten Fall über die Berechnung des euklidischen Abstands berechnet werden:

$$d(x,y) = |x-y| = \sqrt{\sum_{i=1}^{n} w_i (x_i - y_i)^2}$$

Oft werden einige Merkmale wichtiger bewertet werden, als andere. Mit der Einführung von Gewichten (w_i) können Merkmale unterschiedlich skaliert werden.

Um die Klassengrenzen zu visualisieren, bietet sich das Voronoi-Diagramm (Abbildung 3-1) an. In diesem Diagramm werden alle Datenpunkte von einem Polygon begrenzt. Diese legen so ein ganzes

[25] Pruning: Ziel ist es, das Regelwerk möglichst klein und effektiv zu halten. Deshalb kann der Entscheidungsbaum entweder während des Erstellens (prepruning) oder nachträglich (postpruning) gestutzt werden, indem Wurzelknoten/Blätter weggekürzt werden bzw. der Baum komplett umgeformt wird.

Gebiet fest, in dessen Zentrum der Datenpunkt selbst liegt. Derjenige Punkt, der im gleichen Gebiet liegt, ist damit der nearest neighbour.

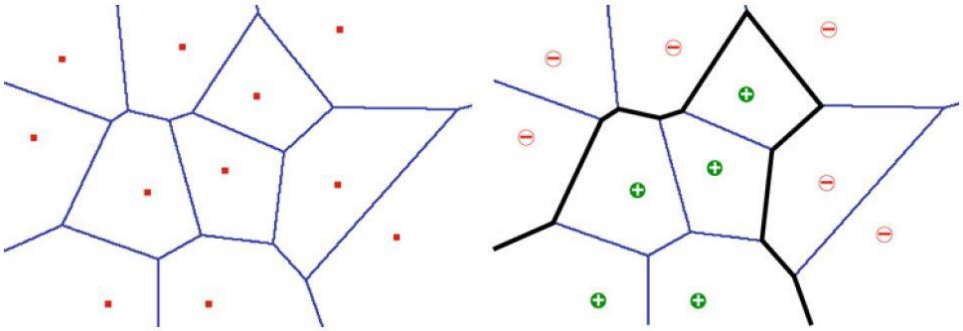


Abbildung 3-1: Eine Punktmenge in einem Voronoi-Diagramm (links) und die daraus resultierende Trennlinie zwischen den beiden Klassen.
(Quelle: Ertel, W.: „Grundkurs Künstliche Intelligenz. Eine praxisorientierte Einführung" (S. 208)).

Da beim Nearest-Neighbour-Verfahren in der Lernphase außer dem Abspeichern der Daten keine weiteren Vorgänge aktiv sind, wird diese Methode auch als „Lazy Learning" bezeichnet. Das Modell wird also nur in diesem spezifischen Moment gebildet, in dem abgefragt wird. Im Gegensatz dazu steht das „Eager Learning", bei dem die Lernphase aufwändig, die Anwendung auf neue Beispiele dadurch aber sehr effizient sein kann. Hier findet die Modellbildung offline statt und kann damit bereits Ergebnisse während der Abfrage liefern. Zur Rubrik des Eager Learning zählen alle Verfahren der neuronalen Netze sowie die Entscheidungsbäume.

3.2 Unüberwachtes Lernen

Beim unüberwachten Lernen sind die Ausgänge nicht bekannt. Der Algorithmus soll vielmehr von selbst Ähnlichkeiten und Muster in den zu analysierenden Daten finden und die Ergebnisse eigenständig kategorisieren.

3.2.1 Assoziative Regeln

Die Anwendung der Assoziationsregeln lässt sich am besten mit der Warenkorbanalyse beschreiben. Durch den Strichcode werden an der Kasse das Produkt, der Hersteller und die Packungsgröße erkannt. Zum einen kann anhand der Artikelnummer automatisch ein Kassenzettel aufgrund der gespeicherten Preistabelle erstellt werden. Als Nebenprodukt kann so das Kundenverhalten analysiert werden. In Datenbanken wird abgelegt, aus welchen Artikeln der Einkauf bestand. Auf Basis dieser Erkenntnisse kann der Supermarkt im weiteren Verlauf seine Produkte optimiert anordnen, um die Verkaufszahlen zu erhöhen.

Es handelt sich hierbei um ein Schema nach dem Muster „Wenn (Item A)…, dann (Item B) …“. Dabei sollen also Items (Elemente einer Menge) ermittelt werden, die das Auftreten anderer Items innerhalb derselben Transaktion implizieren.

Für diese Analyse ist ein Verfahren in zwei Stufen notwendig. Innerhalb dieses Verfahrens werden die beiden Messzahlen, der Support und die Konfidenz, berechnet. Aus diesen beiden Messzahlen werden die Assoziationsregeln entwickelt.

Grundlage bildet die Datenbasis (beispielsweise alle Einkäufe aller Kunden an einem Tag). Mit dem Support wird angegeben, mit welcher Häufigkeit eine Regel anwendbar ist. Beispielsweise wie oft Schuhe zusammen mit Socken eingekauft wurden. Dafür wird die Anzahl der Transaktionen, bei denen die Regel angewendet werden kann, durch die Größe der Datenbasis dividiert. Die Konfidenz gibt daraufhin an, mit welcher Häufigkeit Schuhe und Socken gekauft wurden, wenn alle Transaktionen mit Schuhen berücksichtig werden.

$$confidence(\{Schuhe\} \rightarrow \{Socken\}) = \frac{support\ (\{Schuhe, Socken\})}{support\ (\{Schuhe\})}$$

Der *AIS*[26]*-Algorithmus*[27] ist der erste Algorithmus der die Assoziationsanalyse implementiert. Grundlage des Algorithmus ist der frührere *Apriori-Algorithmus.* Dieser wird jedoch mittlerweile aus Effizienzgründen nicht mehr verwendet.

3.2.2 Clusteranalyse

Bei der Clusteranalyse werden Datenmengen durch die Entdeckung ähnlicher Strukturen analysiert. Dabei werden Gruppen ähnlicher Muster als Cluster bezeichnet. Der Vorgang der Bildung dieser Gruppen wird als „Clustering“ bezeichnet.

Ziel der Clusteranalyse ist es, neue Gruppen in unstrukturierten Daten zu entdecken. Diese Funktionsweise ist gegensätzlich zur Klassifikation, bei welcher Daten bereits bekannten Gruppen zugeordnet werden.

[26] Benannt nach seinen Erfindern Agrawal, Imielinski und Swami

[27] Vgl. Rakesh Agrawal, Tomasz Imielinski, Arun Swami: Database Mining: A Performance Perspective. In: IEEE Trans. Knowl. Data Eng., San Jose 1993, S. 914-925.

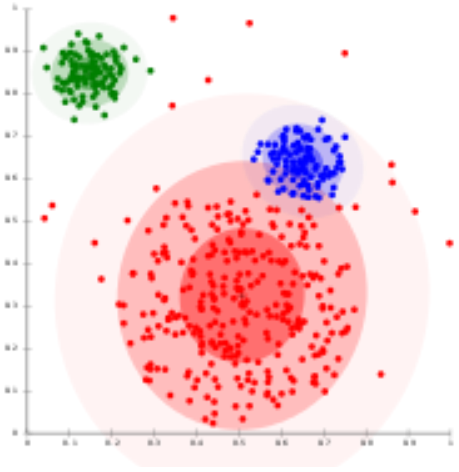


Abbildung 3-2: Ergebnis einer Clusteranalyse mit Normalverteilung.

(Quelle: https://de.wikipedia.org/wiki/Clusteranalyse).

Dieses Verfahren bietet sich dann an, wenn „man statt einer großen Zahl einzelner Instanzen lieber mit einer überschaubar kleinen Anzahl idealerweise homogener Gruppen umgehen möchte."[28] Im Idealfall sollte die Clusteranalyse als Ergebnis nicht nur eine Einteilung der Instanzen in die verschiedenen Gruppen liefern, sondern auch eine leicht verständliche Beschreibung der Gruppen bieten. Das Clustering kann dabei in drei Verfahren unterteilt werden:

1.) *Nichtüberlappende Gruppen*: Hierbei wird jede Instanz nur genau einer Klasse zugeordnet.

2.) *Überlappende Gruppen*: Eine Instanz kann mehreren Gruppen zugeordnet werden.

3.) *Fuzzy-Gruppen*: Die Instanz gehört jeder Gruppe zu einem bestimmten Anteil an.

Verantwortlich für die Zuordnung sind die Clusteralgorithmen. Hier wird zwischen „harten"[29] und „weichen"[30] Methoden unterschieden. Die harten Methoden ordnen jeden Datenpunkt nur genau einem Cluster zu. Die weichen Methoden hingegen ordnen jeden Datenpunkt jedem Cluster zu einem bestimmten Grad zu.

3.3 Reinforcement Learning

Dieser Lernprozess arbeitet mit Verstärkungen in Form von Bestrafungen oder Belohnungen (*rewards*). Diese erhält der Agent während oder am Ende einer Aktion. Bleibt beispielsweise einem Taxiagenten am Ende einer Taxifahrt das Trinkgeld verwehrt, ist dies unter Umständen ein Hinweis auf einen Fehler. Andererseits sind die Punkte am Ende eines gewonnenen Spiels der Beweis für eine gute Leistung. Es ist allerdings Aufgabe des Agenten zu entscheiden, welche Aktion letztendlich die Belohnung oder Bestrafung ausgelöst hat. Beim *Reinforcement Learning* kann ein Agent also durch seine eigene Wahrnehmung und gelegentliche Verstärkung Erfahrungen in einer unbe-

[28] Wrobel, Joachims, Morik: Maschinelles Lernen, S. 457.
[29] Beispielsweise: k-means-Verfahren, Spectral Clustering, Kernel PCA.
[30] Beispielsweise: EM-Algorithmus mit Gauss-Mischverteilung.

kannten Umgebung (Welt) sammeln.[31] Er optimiert also selbstständig eine Nutzenfunktion, die den Wert eines Zustands oder einer Aktion beschreibt.

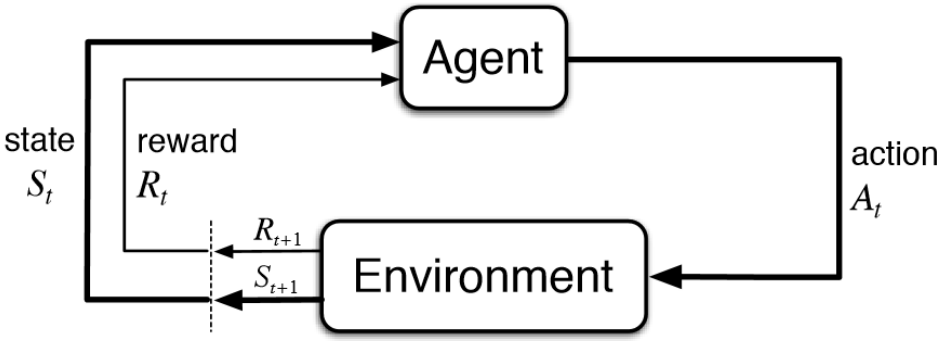


Abbildung 3-3: Beziehung zwischen Agent und Umwelt (Quelle: http://web.stanford.edu/class/cs234/index.html (Stand: 17.11.2017)).

In Abbildung 3-3 ist die Interaktion eines Agenten mit seiner Umwelt dargestellt. Der Agent erhält nach einer Aktion an der Umgebung eine Zustandsbeschreibung und eine Verstärkung in Form einer Belohnung oder Bestrafung. Auf Grundlage seiner bisherigen Erfahrung und diesen Informationen trifft der Agent die Entscheidung für seine nächste Aktion.

„Aufgabe des verstärkenden Lernens ist es, anhand beobachteter Belohnungen eine optimale (oder fast optimale) Strategie für die Umgebung zu lernen".[32] Das Problem kann als Markov-Entscheidungsprozess (MEPs) angesehen werden.

Das Modell des Reinforcement Learning beruht damit auf natürlich menschlichem Lernverhalten, ähnlich Neuronalen Netzen (Kapitel 4).

[31] Vgl. Russell, Norvig: Künstliche Intelligenz, S. 960ff.
[32] Russell, Norvig: Künstliche Intelligenz, S. 960.

4 Neuronale Netze

Künstliche neuronale Netze wurden bereits in den Anfängen der KI-Forschung beschrieben. In dem Artikel „A logical calculus of the ideas immanent in nervous activity" von Pitts und McCulloch aus dem Jahr 1943 wird das mathematische Modell eines Neurons vorgestellt. Dieses Modell sieht Neuronen als grundlegende Schaltelemente im Gehirn vor. Damit war der Grundstein für den Bau künstlicher neuronaler Netze gelegt – einem äußerst wichtigen Teilgebiet der KI.

Dieses System wird, im Gegensatz zu anderen Anwendersystemen, nicht programmiert, sondern trainiert. Dass detaillierte Informationen über die Arbeitsweise des Gehirns immer noch nicht vorliegen, liegt unter anderem an diesen Neuronen. Bis zu 100 Milliarden Neuronen hat der Mensch. Wie in 4.1 erklärt wird, sind sie über Axone und Dendriten miteinander verbunden. Da jedes dieser Neurone mit 1000 bis 10.000 anderen Neuronen ein Netzwerk bildet, ergibt sich eine Gesamtheit von 10^{14} Verbindungen.[33]

In den folgenden Kapiteln werden der Aufbau und die Funktionsweise des Gehirns näher betrachtet, bevor es um die Modelle der künstlichen neuronalen Netze (KNN) und des Deep Learning, einer Weiterentwicklung der KNNs, geht.

4.1 Natürliche Neuronale Netze

Das Gehirn besteht aus etwa 100 Milliarden Nervenzellen (Neuronen), die netzwerkartig miteinander verbunden sind. Jedes Neuron besteht aus einem Zellkörper, einem Zellkern, einem Axon und mehreren Tausend Dendriten (Abbildung 4-1). Der Zellkörper ist für die Speicherung der elektrischen Spannungen verantwortlich: er empfängt Ladung durch die Spannungsimpulse anderer Neuronen und arbeitet ähnlich einem Akku. Des Weiteren besitzt der Zellkörper Verästelungen, sogenannte Dendriten. Eines davon ist das Axon.

Dieses Axon ist der Ausgang des Neurons und kann sich ebenfalls weiter verzweigen. Die Dendriten sind die Eingänge des Neurons. Mit ihrer Hilfe kann das Neuron die elektrischen Impulse empfangen.

[33] Vgl. Wolfgang Ertel: Grundkurs Künstliche Intelligenz. Eine praxisorientierte Einführung. Wiesbaden 2016, S. 265.

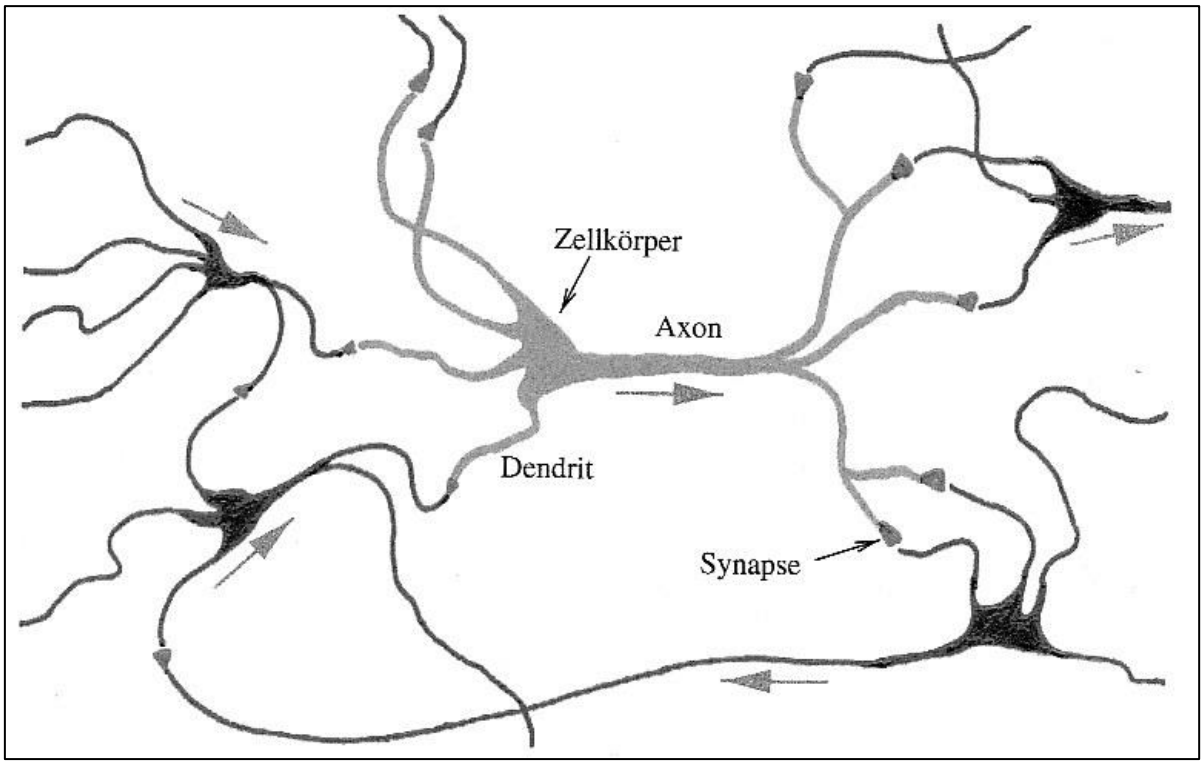


Abbildung 4-1: Modell eines biologischen Neurons im Gehirn
(Quelle: Ertel, W.: „Grundkurs Künstliche Intelligenz. Eine praxisorientierte Einführung" (S. 267)).

Über das Axon können diese Impulse verteilt werden. Dazu „schwebt" das Axon über einem Dendrit eines anderen Neurons. Es gibt keine direkte Berührung zwischen Axon und Dendrit: stattdessen entsteht der sogenannte „Synaptische Spalt". Für Elektronen ist es unmöglich, diesen Spalt zu überwinden. Allerdings ist das Ende eines Axons mit chemischen Botenstoffen (Neurotransmittern) gefüllt. Diese werden von der Spannung ionisiert und können so die Ladung auf den Dendriten des nächsten Neurons übertragen. Dort werden die Neurotransmitter wieder in elektrischen Strom zurückgewandelt und in den Zellkörper weitergeleitet. Die Synapsen sind sozusagen Regler, die selbst festlegen, welche Menge des elektrischen Stroms den Spalt überwinden darf. Der Anteil der Spannung hängt unter anderem von der Konzentration und der chemischen Zusammensetzung der Neurotransmitter ab.

Die Oberflächenspannung eines nicht gereizten Neurons, das sogenannte Ruhepotential, beträgt etwa -70 mV. Durch den ankommenden Spannungsimpuls erhöht sich das Ruhepotential kurzfristig. Die ankommenden Spannungen an allen Dendriten werden addiert. Wird ein festgelegter Schwellwert überschritten, gibt das Neuron über das Axon an den nächsten Dendriten eines Neurons einen Spannungsimpuls ab – es „feuert". Findet dieser Vorgang bei mehreren Neuronen gleichzeitig statt, wird er als „Denken" bzw. „Lernen" beschrieben.

Somit werden die synaptischen Regler bei jedem Lernprozess verändert. „In der Tat ist es so (stark vereinfacht), dass unser Wissen durch die Gesamtheit der Kombinationen aller synaptischer Reglerstellungen in unserem Gehirn repräsentiert wird. Neues Wissen lernen heißt also nicht die Zunahme irgendwelcher Speicherbelegungen, sondern einfach nur die Veränderung der Kombination der synaptischen Regler.“[34] Durch die ständige Veränderung der Synapsen ist es auch zu erklären, dass bereits Gelerntes ganz oder teilweise vergessen wird. Es sind also nicht die Neuronen adaptiv, sondern die Synapsen und damit die Verbindungen zwischen den Neuronen.

4.2 Künstliche Neuronale Netze

Die Idee künstlicher neuronaler Netze (KNN) ist es, Informationen auf verschiedenen synaptischen Reglern abzulegen. Dabei müssen die biologischen Neuronen auf dem Computer nachmodelliert werden.

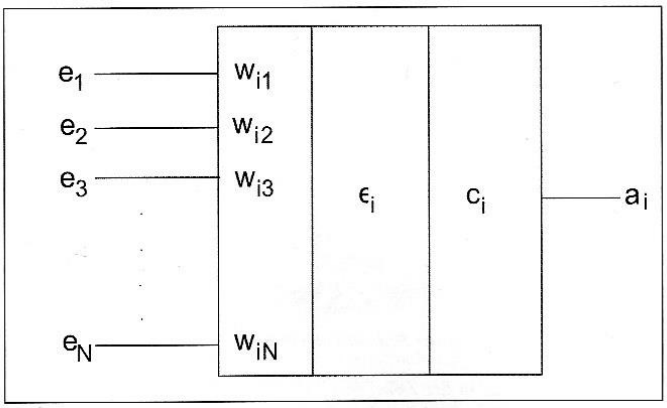


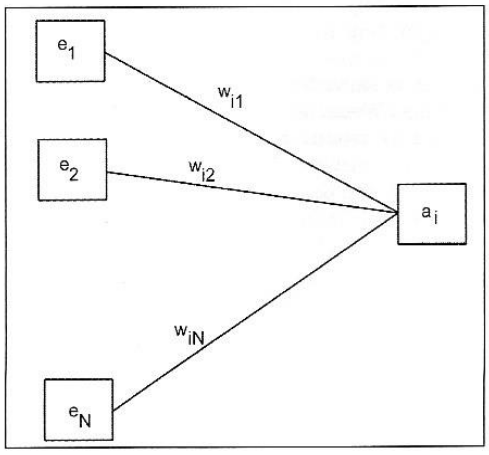
Abbildung 4-2: Schema eines künstlichen Neurons.
(Quelle: Zöllner-Greer, P.: „Künstliche Intelligenz. Grundlagen und Anwendungen" (S. 104)).

Im Blockschaltbild (Abbildung 4-2) wird angenommen, dass dieses Neuron Teil einer Gruppe von Neuronen ist, die alle den gleichen Input e_1 bis e_N eingespeist bekommen. Nach der Schreibweise von Norbert Hoffmann[35] werden nun die einzelnen Variablen erklärt.

[34] Zöller-Greer: Künstliche Intelligenz. Grundlagen und Anwendungen, S. 103 f.
[35] Norbert Hoffman: Kleines Handbuch Neuronale Netze. Wiesbaden 1993.

Symbol	Bedeutung	Biologische Entsprechung
i	i-tes Neuron	i-te Nervenzelle im Gehirn
j	j-ter Eingang des Neurons	j-tes Dendrit der Nervenzelle
e_j	j-ter Eingangswert des Neurons	am j-ten Dendrit ankommende Spannung vor der Synapse
w_{ij}	Gewicht des j-ten Eingangs des i-ten Neurons	Durchlasswert der Synapse, die die ankommende Spannung am j-ten Dendrit der i-ten Nervenzelle regelt
$\varepsilon_i = \varepsilon_i (e_j, w_{ij})$	effektiver Eingang des i-ten Neurons	kumulierte Eingangsspannung aller Eingangswerte nach den Synapsen
$c_i = c_i (\varepsilon_i)$	Aktivität des i-ten Neurons	sich einstellendes Membranpotential an der Zelloberfläche
a_i	Ausgangswert	am Axon ausgegebene Spannung

Tabelle 2: Definition der Variablen – nach Norbert Hoffmann.
(Quelle: Zöller-Greer: „Künstliche Intelligenz. Grundlagen und Anwendungen." (S. 105)).



Als weitere Darstellungsform wird in der Literatur häufig die Rezeptorendarstellung verwendet. Dabei stellen die Rechtecke, im Gegenteil zum Blockschaltbild, jeweils Ein- oder Ausgänge dar und die Kanten repräsentieren die Gewichte w_{ij}.

Abbildung 4-3: Rezeptorendarstellung eines Neurons.

(Quelle: Zöller-Greer: Künstliche Intelligenz. Grundlagen und Anwendungen. (S. 105)).

Eine mathematische Funktion kann auf unterschiedliche Weisen modelliert werden: Es können verschiedene Neuronentypen eingesetzt werden, die wiederum unterschiedliche Funktionen ausdrücken. Obwohl natürlich versucht wird, den realen Verhältnissen eines menschlichen Gehirns nahezukommen, sind noch gar nicht alle Funktionsweisen vollständig bekannt. Und diejenigen, die bekannt sind, sind teilweise schwer mathematisch umsetzbar. Trotzdem haben sich Ansätze herauskristallisiert, die weniger komplex als oft vermutet sind. In der Praxis sind diese oft erfolgreich.

Im Folgenden wird eine Auswahl mathematischer Funktionen zur Modellierung vorgestellt.

Berechnung des effektiven Eingangs:

$$\varepsilon = \sum_{j=1}^{n} w_{ij} e_j.$$

Diese Vorschrift gilt allerdings nur für Neuronen erster Ordnung (lineare Neuronen). Für Neuronen höherer Ordnung fließen alle Kombination mit in die Berechnung ein.

Berechnung der Aktivität:

Im einfachsten Fall ist die Aktivität proportional zum effektiven Eingang (lineare Aktivierungsfunktion):

$$c = s * \varepsilon.$$

Hierbei wird die Zeitabhängigkeit nicht berücksichtig. Anders ist dies bei der *BSB-Aktivierungsfunktion*[36]. In diesem Modell wächst die Oberflächenspannung kontinuierlich an, solange Signale auf die Dendriten treffen. Bleiben die Signale aus, kehrt die Oberflächenspannung wieder zum Ruhepotential zurück.

Die Hopfield[37]-Aktivität stellt ebenfalls eine zeitabhängige Aktivierungsfunktion dar. Dabei werden binäre Neuronen mit den Werten -1 (nicht aktiv) und 1 (aktiv) definiert. Damit hängt die Aktivität nur vom Vorzeichen des effektiven Eingangs ab.

Berechnung des Ausgangs:

Wird der Schwellwert überschritten, „feuert" das Neuron. Abbildung 4-4 visualisiert die unterschiedliche Arten der Modellierung:

a.) Stufenfunktion: Sobald der Schwellenwert der Aktivität überschritten wurde, springt der Funktionswert vom Minimum auf das Maximum.

b.) Sigmoide Funktion: Der Graph in Form einer S-Kurve bietet eine verbesserte Differenzierbarkeit und ist gleichzeitig von oben und unten begrenzt.

[36] BSB ist eine Abkürzung für Brain-State-In-The-Box.
[37] benannt nach John Joseph Hopfield (*1933), amerik. Physiker und Neurowissenschaftler.

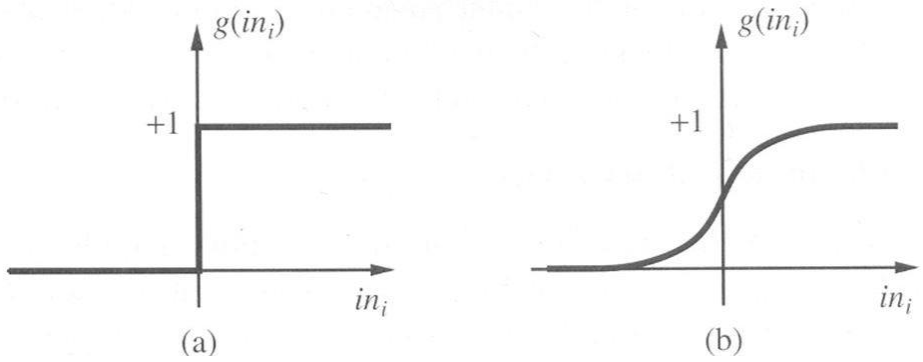


Abbildung 4-4: Stufenfunktion und sigmoide Funktion im Vergleich. (Quelle: http://ki.informatik.uni-wuerzburg.de/teach/documents/KI/skript/Kap20.pdf)

Für die Entwicklung eines neuronalen Netzes sind also von vorneherein viele Entscheidungen zu treffen. Deshalb hat es sich als sinnvoll erwiesen, die Neuronen in Schichten anzuordnen.

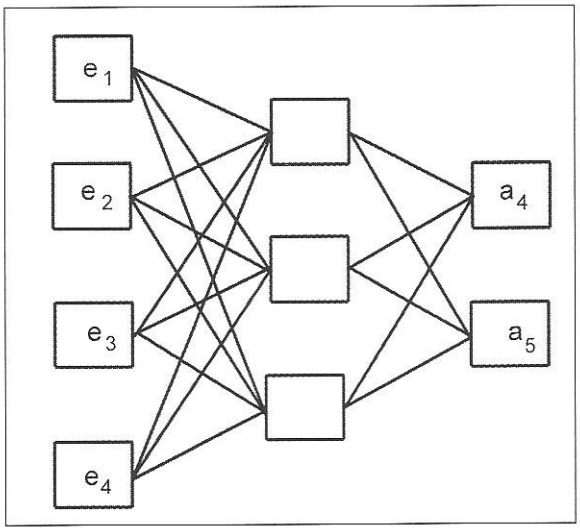


Abbildung 4-5: Rezeptorschichtdarstellung.
(Quelle: Zöller-Greer: „Künstliche Intelligenz. Grundlagen und Anwendungen.“ (S. 113)).

Die Beschreibung der Netze wird unter dem Begriff der Topologie zusammengefasst. Es wird dabei zwischen ein- und mehrschichtigen Netzen unterschieden. Jedes Netz besitzt eine Ein- und eine Ausgangsschicht. Dazwischenliegenden Schichten werden als versteckte Schichten[38] bezeichnet (Abbildung 4-5).

Sind Neuronentyp und Netzwerktopologie entschieden, wird die Art des Lernens festgelegt: Überwachtes oder unüberwachtes Lernen. Erst dann kann die Trainingsphase starten. Hierbei werden die Gewichte justiert, bis die Ein- und Ausgabewerte im vorgegebenen Rahmen übereinstimmen. Darauf folgt die Testphase. „Dort werden dem Netz jetzt Eingaben gegeben, deren zugehörige Ausgaben dem Tester bekannt sind, die aber zuvor noch nie trainiert wurden. Das Netz berechnet seine Ausgaben und es wird dann die Abweichung mit dem tatsächlichen Soll-Output untersucht. Ist

[38] engl. „hidden layer“.

auch diese vom Netz berechnete und nicht trainierte Ausgabe richtig berechnet worden, so gilt die Testphase als abgeschlossen und das Netz kann in die Produktion gehen."[39]

KNNs werden häufig eingesetzt, wenn wenige bis keine strukturierten Daten vorliegen - beispielsweise in der Tex-, Bild- oder Gesichtserkennung. Hier müssen Millionen von Bildpunkten in eine nur geringe Anzahl von möglichen Ereignissen übersetzt werden. Aber auch in der Früherkennung gefährlicher Wetterphänome (beispielsweise von Tornados) oder in der medizinischen Diagnostik werden KNNs eingesetzt.

4.3 Deep Learning

Während der Anfangszeit der KI war es ein Leichtes für einen Computer, Aufgaben zu lösen, die auf mathematischen Regeln basierten. Dagegen stellen diese Aufgaben hohe Anforderungen an einen Menschen. Im genauen Gegenteil liegt jedoch die wahre Herausforderung einer künstlichen Intelligenz: nämlich die Aufgaben zu lösen, die für den Menschen leicht und intuitiv auszuführen sind, wie beispielsweise die Sprach- und Gesichtserkennung.

Forscher der 1950er Jahre blickten bereits in diesem frühen Stadium der KI-Forschung optimistisch in die Zukunft. Allerdings wurde ihnen schnell klar, welche komplexen Aufgaben das menschliche Gehirn regelmäßig vollführt und wie schwierig es ist, diese Leistungen regelkonform abzubilden. Infolgedessen wurden jahrzehntelang Regeln und Merkmale in Computer eingegeben, um dem Computer die Möglichkeit zu geben, Gegenstände anhand ihrer Merkmale zu unterscheiden – ein mühevolles Unterfangen.

Erst in den 1980er Jahren wurde dieses Vorgehen abgelöst. Als neuronale Netze ihren zweiten Frühling erlebten, wurde schließlich eine Möglichkeit entdeckt, die Formulierung der Regeln und Merkmale dem Computer selbst zu überlassen.

Beim *Deep Learning* werden die künstlichen Neuronen in mehrere Ebenen eingeteilt. Dabei ist jede Ebene auf ein Merkmal spezialisiert. Bei Bilderkennung z.B. unterscheidet die erste Ebene Helligkeitswerte, die zweite Ebene erkennt zu Kanten verbundene Pixel (Abbildung 4-6). Die Richtung dieser Kanten kann in der dritten Ebene definiert werden. Somit wird das Bild nach und nach entschlüsselt, wobei jede Ebene auf den Erkenntnissen der Vorherigen aufbaut.

[39] Zöller-Greer: Künstliche Intelligenz. Grundlagen und Anwendungen, S. 115.

Die Besonderheit des Deep Learning liegt in der Mehrschichtigkeit der neuronalen Netzwerke. Das Netzwerk ist tiefer, je mehr Ebenen definiert werden. Der Abstraktionsgrad der Information nimmt mit der Entfernung der Schicht von der Eingabeschicht zu. [40]

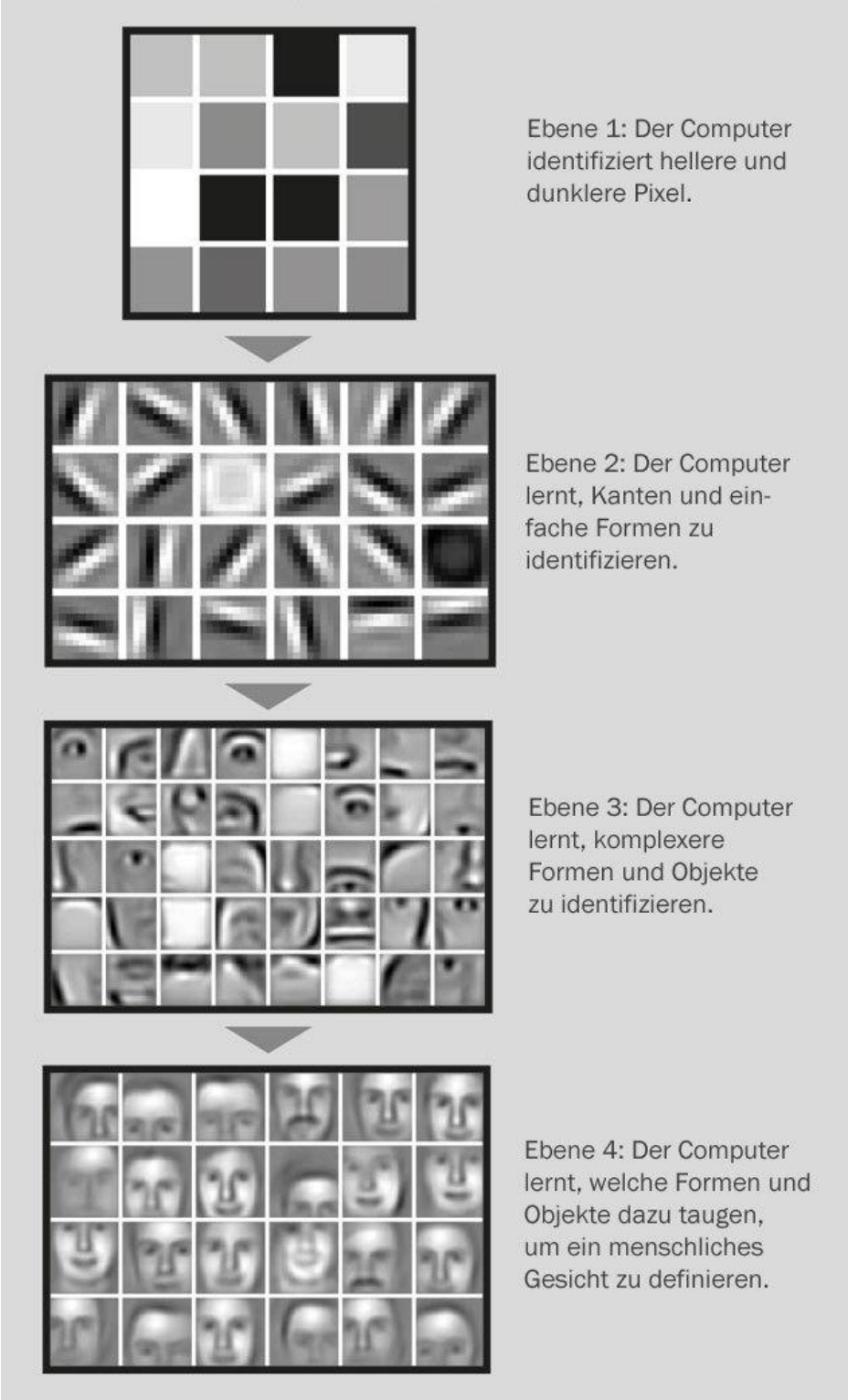


Abbildung 4-6: Gesichtserkennung via Deep Learning
(Quelle: Jones, N.: „The learning machines". In: Nature 505, London 2014, S. 146f).

[40] Vgl. Nicola Jones: The learning machines. In: Nature 505, London 2014, S. 146-148.

5 Spracherkennung

Sprache und Denken sind eng miteinander verbundene Bereiche der menschlichen Intelligenz. „Spracherwerb kommt einerseits ohne ein Minimum an kognitiven Fähigkeiten nicht aus, ist andererseits aber auch Vorbedingung für deutliche Fortschritte in der allgemeinen Intelligenzentwicklung."[41] Damit ist die Sprache nicht nur eine zwingende Voraussetzung für soziale Interaktion, allgemeinen Erkenntnisgewinn und Wissenstransfer. Ohne ausreichende Sprachkenntnisse sind auch die Fähigkeiten Kategorisierung, Benennung und Abstraktion nur schwer möglich.

Es ist nicht verwunderlich, dass die Spracherkennung von Beginn an eine zentrale Rolle im Bereich der KI gespielt hat. Die Aufnahme eines Sprachsignals mit Mikrofon und die anschließende Digitalisierung sind heutzutage kein Problem mehr. Im Bereich der Spracherkennung wird das Signal als Eingangsdaten verwendet; zur Mustererkennung wird es in eine Folge von Symbolen übersetzt.

Wichtig ist die Unterscheidung zwischen sprecherunabhängiger und sprecherabhängiger Erkennung. Im Fall einer „sprecherunabhängigen Erkennung erfolgt keine Adaption des Spracherkennungssystems auf die Artikulation einzelner Sprecher, während bei der sprecherabhängigen Erkennung vor der ersten Nutzung ein Trainingslauf durchzuführen ist, bei dem durch Vorlesen eines vorgegeben Textes Besonderheiten der Aussprache einzelner Nutzer erfasst werden, um später bessere Erkennungsergebnisse erzielen zu können. Durch dieses Training können auch erheblich größere Vokabularien als bei der sprecherunabhängigen Erkennung zum Einsatz kommen."[42]

Einen Mittelweg bietet die sprecheradaptive Erkennung: Dabei passt sich, ausgehend von der sprecherunabhängigen Erkennung das System an individuelle Eigenheiten des Sprechers an. Weitere Aufgabenstellungen können die Sprechererkennung (engl. Speaker-Indexing) und Sprachenerkennung sein. Beim Speaker-Indexing werden Datenbanken nach bestimmten Sprechern durchsucht, um schnell bestimmte O-Töne zu finden. Ein ähnlicher Anwendungsbereich ist die automatisierte Erkennung von Schlagwörtern (keywords) um archiviertes Audiomaterial schnell durchsuchen zu können, falls keine Transkription zur Verfügung steht.[43]

---

[41] Wolfgang Menzel: Sprachverarbeitung. In: Görz, G., Schneeberger, J., Schmid, U.: Handbuch der Künstlichen Intelligenz. Oldenburg 2014, S. 473.

[42] Bernd Ludwig: Spracherkennung. In: Kuhlen, R., Semar, W., Strauch, D.: Grundlagen der praktischen Information und Dokumentation. Berlin 2013, S. 313.

[43] Vgl. Stephen Euler: Grundkurs Spracherkennung. Wiesbaden 2006, S. 15.

Die häufigsten Anwendungsbereiche der Spracherkennung sind die Gerätesteuerung, Diktier- und Sprachdialogsysteme.

5.1 Sprachverarbeitung

5.1.1 Schwierigkeiten in der Sprachverarbeitung

Bei natürlicher Sprache handelt es sich – im Gegensatz zu formalen Programmiersprachen, die auf wenige Grundregeln begrenzt sind – um eine „lebendige" Sprache. Sie befindet sich in einem dauerhaften Entwicklungsprozess: Stetige Veränderungen wirken sich auf den Wortschatz, die Grammatik und Betonung aus. Gleichzeitig verschwinden veraltete Wörter komplett aus dem Sprachgebrauch. Zudem ist die Komplexität der Grammatik sprachenabhängig. Gerade im Deutschen gibt es viele Sonderfälle und Ausnahmen. Die Flexion und Trennung des Verbs ist dabei von zentraler Bedeutung. Die semantische Unterteilung des Verbs „lieben" wird beispielhaft in Tabelle 3 dargestellt.

Wurzel	**(lieb)**
Flexion	(((lieb)**t)en)**
Präfigierung	**(ver**(lieb))
Suffigierung	((lieb)**lich)**
Zirkumfigierung	**(ge**(lieb)**t)**

Tabelle 3: Elementare Bausteine in der Wortzusammensetzung.
(Quelle: Menzel, W.: „Sprachverarbeitung". In: Görz, G., Schneeberger, J., Schmid, U.: „Handbuch der Künstlichen Intelligenz" (S. 487)).

Außerdem sind oft weitere Informationen (Kontextwissen) notwendig, um den Satz eindeutig zu verstehen. Die folgenden Beispiele verdeutlichen dies:

1.) Das Mädchen ist verzogen.

Dieser Satz birgt zwei Verständnismöglichkeiten. Nur aufgrund dieses Satzes lässt sich nicht entscheiden, ob das Mädchen schlecht erzogen (Verb + Adjektiv/Präsens) wurde, oder ob es umgezogen ist (Hilfsverb + Partizip/Perfekt).

2.) Treffen wir uns vor dem Museum?

Mehrdeutigkeit aufgrund des Wortes „vor". Dieses kann sich sowohl auf den Ort beziehen, als auch auf den Zeitpunkt.

3.) Bernd hat einen Hund. Er ist sehr groß.

Hier kann das Pronomen „er“ nicht eindeutig zugeordnet werden. Es kann sich auf den Hund, oder aber auf Bernd beziehen.[44]

4.) Wir müssen noch einen Termin ausmachen.[45]

In geschriebener Formulierung ist es nicht sicher, ob es sich um einen weiteren Termin oder um einen „ersten“ Termin handelt, der hier ausgemacht werden muss. Im alltäglichen Sprachgebrauch lässt sich aufgrund der Betonung der Sinn erschließen.

Des Weiteren gibt es im Deutschen einige Wörter mit mehreren Bedeutungen. Beispielsweise:

- Bank (Geldinstitut oder zum Sitzen)
- Bremse (Insekt oder im Fahrzeug)
- Läufer (Sportler oder Teppich)

Meistens erschließt sich die Bedeutung des Wortes aus dem Kontext. Hinzu kommen die Homophone. Hierunter versteht man Wörter, die gleich klingen und ausgesprochen werden, aber eine unterschiedliche Bedeutung haben (Beispiel: Meer <-> mehr). Diese kann der Computer aufgrund fehlenden semantischen Wissens nicht unterscheiden.

Weitere Probleme im Bereich der Spracherkennung ergeben sich aus der individuellen Sprechweise eines jeden Menschen. Wie in 5.2.1 näher erläutert wird, hat das Sprechen bestimmte Bewegungsmuster, die mit den Bewegungsmustern des Gehens vergleichbar sind. Das heißt, dass die anatomischen Voraussetzungen und die erlernte Technik großen Einfluss auf die individuelle Art des Sprechens haben. Trotzdem „verbleiben nicht kontrollierbare, zufällige Variationen des Artikulationsprozesses. Das komplizierte Zusammenspiel vieler Muskeln lässt sich nicht identisch wiederholen.“[46]

Als weitere Konsequenz daraus ist es unmöglich, sprachliche Ausdrücke exakt zu reproduzieren.[47] Große Abweichungen können auch durch starke Dialekte /Akzente hervorgerufen werden.

---

[44] Vgl. Euler: Spracherkennung, S. 2.
[45] Wahlster, W.: Verbmobil: Foundations of Speech-to-Speech Translation. Berlin 2000, S. 467.
[46] Euler: Grundkurs Spracherkennung, S. 13.
[47] Für weitere Informationen: Fred Englert, Stephen Euler, Dietrich Wolf: Zur Variabilität sprachlicher Äußerungen in der sprecherunabhängigen Einzelworterkennung. In: Mangold, H.: Sprachliche Mensch-Maschine-

Einen kritischen Einfluss auf die Erkennungssicherheit hat die akustische Umgebung. Die Qualitätsunterschiede (beispielsweise durch starke Hintergrundgeräusche) während der Aufnahme haben großen Einfluss auf die spätere Spracherkennung.

Hieraus folgt, dass verschiedene Anwendungsfälle unterschiedliche Herausforderungen an das Spracherkennungssystem stellen (Tabelle 4).

	Einfach	**Schwierig**
Sprechweise	einzelne Wörter	Sätze
Wortschatz	kleiner Wortschatz	großer Wortschatz
Benutzerkreis	sprecherabhängig	sprecherunabhängig

Tabelle 4: Unterscheidungskriterien für Spracherkennung.
(Quelle: Euler, S.: „Grundkurs Spracherkennung“ (S. 18)).

5.1.2 Aufbau von Spracherkennungssystemen

Um diesen Anforderungen gerecht zu werden, hat sich das folgende Modell (Abbildung 5-1) der Spracherkennung bewährt.

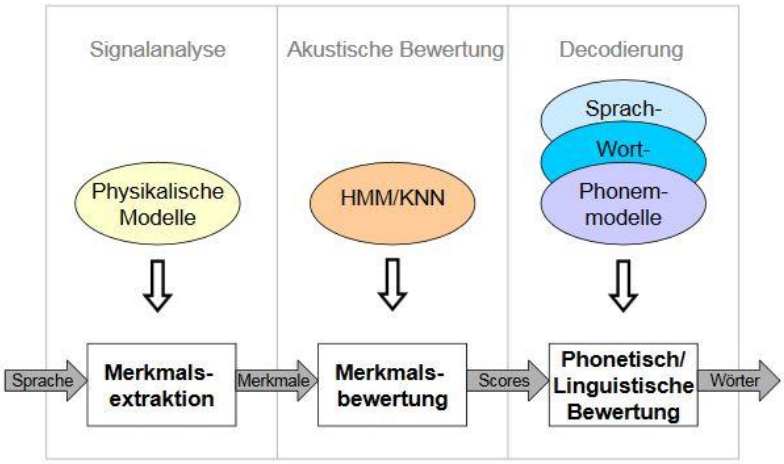


Abbildung 5-1: Prinzipieller Aufbau von Spracherkennern.
(Quelle: https://www.fbi.h-da.de/fileadmin/personal/k.kasper/Materialien/ASR/asr3.pdf (Stand: 04.11.2017)).

Diese Aufteilung gibt die groben Einteilungsmöglichkeiten in Bezug auf die Teilabschnitte der Spracherkennung wieder. Allerdings ist eine strikte Trennung der Stufen oft nicht möglich. Vielmehr spielt die Interaktion zwischen den Stufen eine wichtige Rolle.

Kommunikation. München 1992, S. 15 ff. Hier wurde bei einer Auswahl von 100 Äußerungen des Wortes „sieben“ von verschiedenen Sprechern und Sprecherinnen insgesamt 10 unterschiedlichen Aussprachevarianten beobachtet.

Bevor es um die technische Umsetzung der Spracherkennung geht, ist ein kurzer Blick in die Sprachproduktion unverzichtbar.

5.2 Akustische Modelle für gesprochene Sprache

Sprache gilt als das akustische Produkt des menschlichen Sprechapparats. Deshalb werden in diesem Kapitel die Grundlagen der anatomischen und physiologischen Gegebenheiten vorgestellt.

5.2.1 Aufbau und Funktion des Sprechapparats

Zur besseren Unterscheidung lassen sich die Organe in drei Bereiche einteilen (Tabelle 5).

Bereich	Beteiligte Organe	Funktion
supra-laryngales System	Rachen-, Nasen- und Mundraum	Bildung der Laute (Artikulation)
laryngales System	Kehlkopf, Stimmlippen	Stimmgebung (Phonation)
sub-glottales System	Luftröhre, Bronchien, Lunge	Atmung

Tabelle 5: Übersicht der an der Sprachproduktion beteiligten Organe und deren Funktion.

5.2.1.1 Sub-glottales System

Alle Laute des Deutschen (sowie der überwiegende Teil aller Laute in den Sprachen der Welt) werden während des Ausatemvorgangs gebildet. Man spricht dabei von einem egressiven[48] pulmonalen[49] Luftstrom, der von der Lunge durch den Kehlkopf und Vokaltrakt nach außen geführt wird. Das maximale Lungenvolumen beträgt sechs bis sieben Liter. In der Ruhelage ist die Lunge mit drei Litern gefüllt. Während des Einatmens werden etwa 0,5 Liter zusätzlich eingeatmet. Für diesen Vorgang ist Muskelaktivität erforderlich. Das Ausatmen dagegen ist ein passiver Ablauf, der über elastische Kräfte erfolgt.

Ganz gleich welche Laute gebildet werden, die Lunge hält den Luftdruck annähernd konstant. Dies ist besonders wichtig, weil der Schalldruckpegel des Sprachsignals annähernd im Quadrat mit dem sub-glottalen Luftdruck ansteigt, d.h. dass eine Verdopplung des sub-glottalen Luftdrucks zu einer Vervierfachung des Schalldruckpegels führt.[50] Die Regulierung des Luftdrucks führt also gleichzei-

[48] Egressiv <lat.> egressivus „fortsetzen“.
[49] Pulmonal <lat.> pulmo „Lunge“.
[50] Vgl. van den Berg, J.: Direct and indirect determination of mean subglottic pressure. Basel 1956, S. 15f und Williard R. Zmelin: Speech and Hearing Science – Anatomy and Physiology. Boston 1998, S. 158 ff.

tig zur Regulierung der Sprechlautstärke. Soll lauter gesprochen werden, wird meist der Luftdruck erhöht und danach (relativ) konstant gehalten.[51]

5.2.1.2 Laryngales System

Im Kehlkopf befinden sich die Stimmlippen. Diese schwingen bei stimmhaften Sprachlauten mit. Schwingen die Stimmlippen nicht mit, nennt man die Sprachlaute stimmlos. Dieses Oszillieren beeinflusst die Tonhöhe der Sprache.

Beim Sprechen öffnet und schließt sich eine Stimmlippe zwischen 100- und 400-mal pro Sekunde. Beim Singen wird diese Zahl noch übertroffen und ein schreiendes Baby kann die Stimmlippe bis zu 2000-mal pro Sekunde öffnen und schließen. Für diese schnellen Abläufe können also keine Muskel- oder Nervenimpulse verantwortlich sein. Stattdessen lässt sich mittels des Bernoulli-Effekts[52] die Öffnung der zuvor geschlossenen Stimmlippen beschreiben. Demnach entsteht an den engen Stellen des Kehlkopfs eine höhere Strömungsgeschwindigkeit, während der Schalldruck gleichzeitig absinkt. Durch diesen Unterdruck zwischen den Stimmlippen entsteht ein Sog, der diese zusammenzieht und nicht auseinanderdrückt. Dieses hydrodynamische Paradoxon wird auch in der Luftfahrt ausgenutzt: ohne diesen Effekt könnte sich kein Flugzeug in der Luft halten.

Einen wichtigen Einfluss auf die Sprachgrundfrequenz hat die Länge der Stimmlippen. Daraus resultiert bei Männern eine Grundfrequenz von etwa 100 Hz und bei Frauen von etwa 180 Hz.[53]

Vier Komponenten sind für die Schwingungsfrequenz der Stimmlippen verantwortlich:

- Lange Stimmlippen schwingen langsamer als kurze. Der Unterschied in der Länge der Stimmlippen ist anatomisch bedingt.
- Stärker gespannte Stimmlippen schwingen schneller als weniger stark gespannte. Die Spannung wird zur Feinsteuerung der Schwingungsfrequenz beim normalen Sprechen benutzt.
- Stimmlippen mit mehr Masse schwingen langsamer als dünnere.
- Durch den Bernoulli-Effekt schließen sich die Stimmlippen wieder und können so in den Normalzustand zurückgehen. Die Luftdruckunterschiede zwischen dem sub-glottalen und dem supra-laryngalen System haben ebenfalls Einfluss auf die Frequenz.

---

[51] Vgl. Reetz: Artikulatorische und akustische Phonetik. Trier 2003, S. 109.
[52] Daniel Bernoulli (1700-1782), schweiz. Mathematiker und Physiker.
[53] Vgl. Euler: Grundkurs Spracherkennung, S. 5.

Dieses an den Stimmlippen erzeugte Signal ist das Ausgangsmaterial für alle stimmhaften Laute. Sie werden im supra-laryngalen System weiter gefiltert, bevor sie von Lippe und/oder Nase abgestrahlt werden.[54]

5.2.1.3 Supra-laryngales System (Vokaltrakt)

Während beim normalen Atmen die Luft beinahe unbemerkt durch den Nasen- und Rachenraum fließt, werden beim Sprechen Geräusche erzeugt, die als Sprachlaute wahrgenommen werden.

Für die Formung der *Vokale* strömt das Schallsignal in den Vokaltrakt. Die Resonanzen bilden sich durch die Positionen von Zunge und Lippe. Ebenso spielt das weiche Gewebe (u.a. Mundschleimhaut) eine wichtige Rolle: Dadurch wird das Schallsignal in allen Frequenzbereichen, bis auf die Resonanzfrequenz abgedämpft.

Bei der Bildung der *Konsonanten* verengt sich der Vokaltrakt, damit der Laut erzeugt werden kann. Die Luft kann also nicht ungehindert strömen, sondern wird behindert oder bei Verschlusslauten sogar gestoppt.

„In jedem Fall werden Sprachlaute nie direkt in ihrer endgültigen Form erzeugt; es ist immer eine Kombination aus Erzeugung und Veränderung des Signals.“[55] Abhängig von den anatomischen Voraussetzungen und einer erlernten Technik, hat also jeder Mensch eine unterschiedliche Sprechweise. Schlussendlich sind diese Faktoren gemeinsam auch dafür verantwortlich, dass sprachliche Äußerungen nicht exakt nachgeahmt werden können.

5.2.2 Das Quelle-Filter-Modell

Das soeben vorgestellte akustische Modell der Sprachproduktion hat Gunar Fant im Jahr 1960 stark vereinfacht. Nach seinem Quelle-Filter-Modell entsteht die Schallerzeugung aus zwei unabhängigen Teilmechanismen:

1.) Rohschallerzeugung (= Quelle)

2.) Resonator (= Filter)

Demnach durchläuft ein von der Quelle erzeugtes Schallsignal einen Resonator mit einem bestimmten Spektrum bis zur Abstrahlung an den Lippen. Mathematisch gesehen wird das Signal auf diesem Weg mit einer Übertragungsfunktion verformt. Da es sich hierbei um ein lineares System

[54] Vgl. Reetz: Artikulatorische und akustische Phonetik, S. 127 f.
[55] Reetz: Artikulatorische und akustische Phonetik, S. 128.

handelt, kann das Spektrum des abgestrahlten Sprachsignals durch Multiplikation des Quellenspektrums mit der Übertragungsfunktion berechnet werden.[56]

5.2.3 Einheiten

Sprachliche Äußerungen (geschrieben oder gesprochen) lassen sich in kleinere Einheiten unterteilen. Die Morphologie ist die Wissenschaft, die sich mit diesen Einteilungen beschäftigt.

Dabei ist der „Satz" die größte mögliche Einheit und wird durch ein Satzzeichen (Punkt) beendet. In gesprochener Sprache wird ein Satzende durch eine Intonationsänderung und eine Pause verdeutlicht.

Außerdem haben Wörter einen inneren Aufbau. Sie können aus mehreren Wörtern zusammengesetzt sein (z. B. Hausaufgabe) und durch Vor- oder Nachsilben in ihrer Bedeutung verändert werden (z. B. ge-laufen, ver-laufen). Die Worte lassen sich wiederum in (Sprech-) Silben unterteilen. Diese bestehen aus einem Vokal oder Diphthong (Doppelvokal) als Kern, sowie Konsonanten davor oder danach. Diese Sprechsilben[57] sind unabhängig von der vorausgehenden oder nachfolgenden Sprechsilbe.

Als Phon[58] (Laut) wird die kleinste Einheit bezeichnet. Phone werden in der Lautschrift mit speziellen Symbolen dargestellt.[59] Phoneme sind eine Gruppe von Lauten, die ähnlich klingen und niemals einen Bedeutungsunterschied bewirken. Beispielsweise führt die Benutzung des Zungen- oder Zäpfchen-r zu keinem inhaltlichen Unterschied, jedoch handelt es sich um zwei unterschiedliche Phone ([r] und [R]). Die Phone, die gleichzeitig auch Phoneme sind, werden als Allophone bezeichnet.

Da nicht nur der Laut selbst sondern auch dessen Dauer einen Bedeutungsunterschied hervorrufen kann, hängt von der Einteilung der Phoneme auch der Wortschatz ab. Es zeigt sich also, dass die Definition der Lauteinheiten schwierig ist. In Tabelle 6 ist eine Zusammenstellung der sprachlichen Einheiten zu finden. Die Größe des Inventars bezieht sich auf die deutsche Sprache.

---

[56] Vgl. Gunnar Fant: Acoustic Theory of Speech Production. Den Haag 1960.
[57] Im Unterschied zu Schreisilben, die nur bei der Worttrennung eine Rolle spielen.
[58] Phon <griech.> phon „Stimme", „Klang".
[59] Siehe Anhang 1.

Einheit	Inventar (D)
Satz	Unbegrenzt
Wort	Ca. 500.000
Silbe	5000
Phonem	< 50

Tabelle 6: Übersicht der sprachlichen Einheiten. (Quelle: Euler, S.: „Grundkurs Spracherkennung“ (S. 10)).

5.3 Signalanalyse

Zu Beginn des Spracherkennungsprozesses ist es wichtig erkennbar zu machen, wann die Spracheingabe beginnt und endet. Das am Häufigsten angewandte Verfahren basiert auf der Lautstärke des Signals. Die Sprache hebt sich im Normalfall deutlich von den Hintergrundgeräuschen ab. Beginn (t_a) und Ende (t_e) eines Sprachsignals können also anhand der Stärke des Signals s(t) bestimmt werden. Die Stärke *e*, die Energie als Summe der Quadrate über einen kurzen Signalabschnitt (meist ca. 10 ms), lässt sich berechnen mit:

$$e = \sum_{t=t_a}^{t_e} s(t)^2$$

Werden die Werte logarithmiert, können diese besser miteinander verglichen werden:

$$e_l = log_{10} e$$

Um festzustellen, ob es sich um Sprache handelt wird ein Schwellwert (Threshold) eingeführt. Die berechneten Energiewerte werden mit diesem verglichen. Ist die Schwelle überschritten, wird auf Sprache entschieden. Wichtig dabei ist, dass die Energiewerte mehrerer aufeinanderfolgender Blöcke verglichen werden, wobei die Blockgrenzen beispielsweise um 5 ms verschoben werden, um eine Überlappung sicherzustellen. Da sich auch die Hintergrundgeräusche ändern können, ist eine stetige Anpassung sinnvoll. Probleme treten durch stimmlose An- und Auslaute, kurze Störgeräusche, sowie längere Pausen innerhalb des Sprachsignals auf.

Das Sprachsignal kann also näherungsweise als eine Folge von kurzen Blöcken angesehen werden. Zwar ist deshalb nur eine grobe Modellierung möglich, dennoch wird dieses Verfahren bevorzugt verwendet, da hier die Fehlerrate noch minimal ist. Die Auswirkungen von Fehlern in frühen Phasen der Spracherkennung sind später nur schwer zu korrigieren.

5.3.1 Diskrete Fourier-Transformation

Um die Blöcke besser miteinander vergleichen zu können, ist nicht das Signal im Zeitbereich relevant, sondern das Signal im Frequenzbereich. Der Anteil jeder Frequenz wird dann als Kenngröße

verwendet. Deshalb wird eine diskrete Fourier-Transformation (DFT) des Signalblocks durchgeführt. Ist die Blocklänge der DFT gleich der Fensterlänge L, gilt:

$$S(n) = DFT\left(s(l)\right)$$

$$= \sum_{l=t_s}^{t_s+L-1} s(l) * e^{\frac{-2\pi j*n}{L*(l-t_s)}}$$

Dadurch ist das resultierende Frequenzspektrum periodisch mit L. Durch die Entwicklung der *Fast Fourier-Transformation* (FFT) hat sich ein noch effizienteres Verfahren herausgebildet.

Um einen Signalsprung an den Grenzen der Blöcke zu verhindern[60] (dieser könnte zu einer Verfälschung der Frequenzanalyse durch hochfrequente Anteile führen), wird das Signal an den Blockgrenzen abgeschwächt. Dafür wird oft die Fensterfunktion nach Hamming[61] verwendet:

$$h(i) = 0{,}54 - 0{,}46 * \cos\left(\frac{2\pi i}{L}\right) \qquad 0 \leq \mathrm{i} \leq \mathrm{L} - 1.$$

Da die volle Frequenzauflösung für die weitere Verarbeitung zu viele Informationen enthält, werden die Frequenzen in Bänder zusammengefasst. In diesem Fall bietet sich eine Übertragung auf eine Skala an, die sich dem menschlichen Gehör annähert: beispielsweise die Bark[62]-Skala oder die Mel-Skala.

In psychoakustischen Experimenten wurde festgestellt, dass sich die Pegel der einzelnen Frequenzen innerhalb eines Frequenzbandes addieren und letztlich als Gesamtreiz erfasst werden (Bark-Skala). Der Umfang der Frequenzbänder ist frequenzabhängig, wobei die Größe zu hohen Frequenzen zunimmt. Alternativ steht die Mel-Skala zur Verfügung. Dieser Skala liegt ein Verfahren zugrunde, das auf Wahrnehmungsexperimenten beruht. Dabei wurde gemessen, welcher Ton als doppelt so hoch/tief wie eine Referenz wahrgenommen wurde. Es stellte sich heraus, dass bei höheren Frequenzen ein nichtlinearer Zusammenhang besteht.

[60] sogenannter Leck-Effekt.
[61] benannt nach Richard Hamming (1915-1998): amerik. Mathematiker.
[62] benannt nach Heinrich Barkhausen (1881-1956): dt. Physiker.

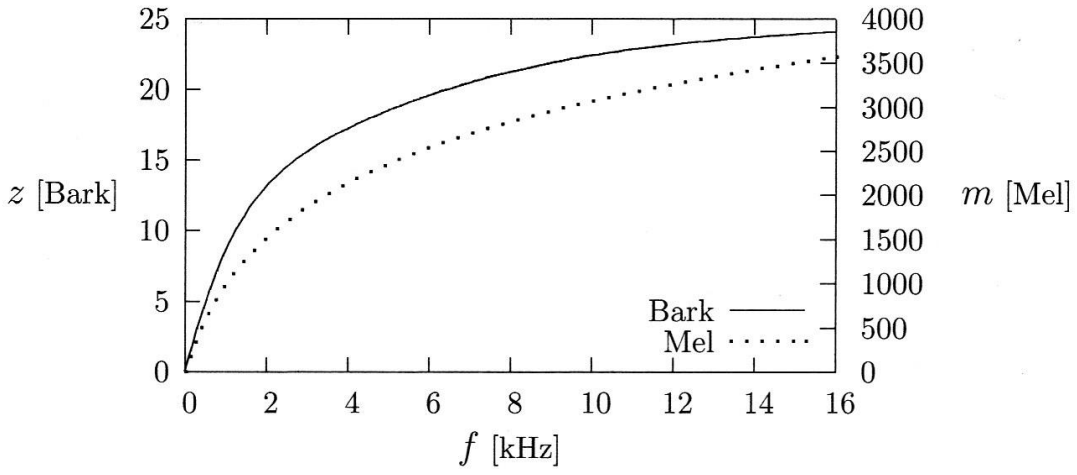


Abbildung 5-2: Zuordnung der linearen Frequenzskalen zur Bark- und Mel-Skala (Quelle: Euler, S.: "Grundkurs Spracherkennung" (S. 42)).

Im Vergleich verlaufen die beiden Skalen bis 500 Hz linear und werden dann zunehmend flacher. Um dem menschlichen Höreindruck näher zu kommen, wurde für den Bereich < 8 kHz eine Mischung aus Bark- und Mel-Skala eingeführt:

< 1 kHz: Abstand zwischen den Mittenfrequenzen liegt konstant bei 100 Hz

> 1 kHz: Oktaven werden logarithmisch in fünf Bänder aufgeteilt

Dadurch wird die Gesamtanzahl der Merkmale auf 20 bis 30 verringert.

5.3.2 Lineare Prädiktive Codierung (LPC)

Die Lineare Prädiktive Codierung (LPC), kurz Lineare Prädiktion, basiert auf dem Quelle-Filter-Modell (siehe 5.2.2) und kann als Alternative zur Merkmalsberechnung angesehen werden. Dieses Modell schließt vom Sprachsignal zurück auf das verwendete Filter. Die Koeffizienten des (abgeleiteten) Filters sind dabei die Merkmale. In diesem Verfahren soll der nächste Signalwert s‘(n) aufgrund der bereits bekannten Werte (s(n-k)) vorhergesehen werden. Für s‘(n) gilt:

$$s'(n) = \sum_{k=1}^{p} a(k) * s(n-k).$$

Als Prädiktorkoeffizienten bezeichnet man die Gewichte a(k). Um diese zu bestimmen, wird die Differenz zwischen dem tatsächlichen Wert und dem Schätzwert, der sogenannte Prädiktionsfehler e(n) berechnet:

$$e(n) = s(n) - s'(n) = s(n) - \sum_{k=1}^{p} a(k) * s(n-k).$$

Mithilfe der Autokorrelationsfunktion (AKF) und durch Verwenden der Symmetrie entsteht beim partiellen Ableiten ein Gleichungssystem zur Bestimmung der Prädiktorkoeffizienten. Mit den Vektoren a und r und der Korrelationsmatrix R lässt sich das Gleichungssystem als

$$R * a = r$$

beschreiben. Zur Lösung dieses Gleichungssystems wird verbreitet der rekursive Levinson-Durbin-Algorithmus verwendet. Die Lösung enthält diejenigen Werte, die die Differenz zwischen Schätzung und Original minimieren.

Wird der Prädiktor als digitales Filter betrachtet, kann die z-Transformation zur Berechnung des Prädiktionsfehlerfilters herangezogen werden. Für den Prädiktionsfehler gilt dann:

$$E(z) = \left[1 - \sum_{k=1}^{p} a(k) * z^{-k}\right] * S(z).$$

Es gilt für den Prädiktionsfehlerfilter:

$$A(z) = 1 - \sum_{k=1}^{p} a(k) * \ z^{-k}$$

Mit dem korrelierten Signal S(z) und dem unkorrelierten Signal E(z):

$$E(z) = A(z) * S(z).$$

Aus dem Fehlersignal kann das ursprüngliche Signal wiederhergestellt werden:

$$S(z) = \frac{1}{A(z)} * E(z).$$

Damit kann das Filter H(z) = 1/A(z) im Quelle-Filter-Modell zur Synthese des Sprachsignals dienen. Durch weiterführende Berechnungen lässt sich zeigen, dass bei ausreichend großer Prädiktorordnung jedes Spektrum mit beliebiger Genauigkeit angenähert werden kann.

Die unterschiedlichen Ergebnisse der linearen Prädiktion und der spektralen Schätzung werden in Abbildung 5-3 gegenübergestellt. Es zeigt die deutliche Annäherung der LPC-Kurve an das Frequenzspektrum nach Fourier.

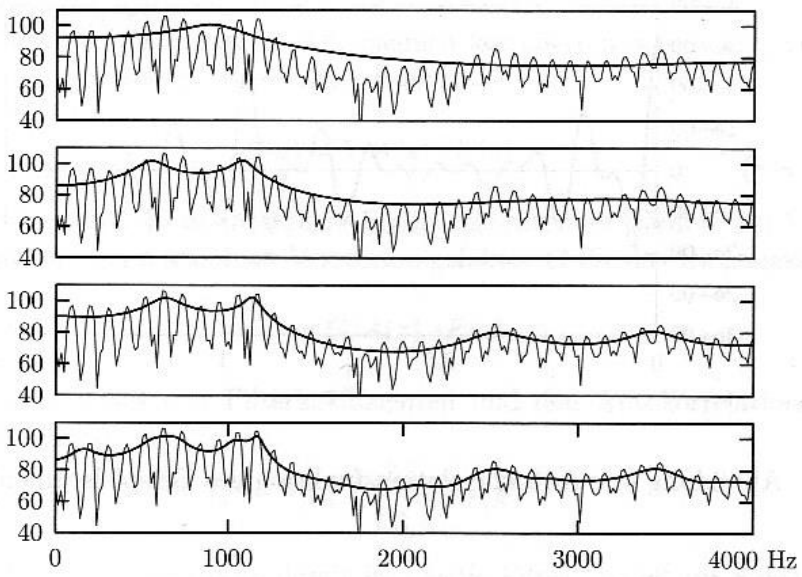


Abbildung 5-3: Approximation eines Sprachspektrums durch Prädiktoren der Ordnung 4, 8, 16 und 32 (mit abfallender Ordnung angeordnet). (Quelle: Euler, S.: "Grundkurs Spracherkennung" (S. 42)).

Mit zunehmender Ordnung der linearen Prädiktion ist gleichzeitig eine Glättung zu Erkennen. Die entscheidenden Verlaufsänderungen sind sichtbar. Lediglich die Details und Vielfachen der Sprachgrundfrequenz gehen verloren. Dies ist allerdings für ein sprecherunabhängiges System durchaus wünschenswert, da das Hauptaugenmerk ohnehin auf einem groben Verlauf liegt.

Die direkte Verbindung zum Quelle-Filter-Modell liegt auf der Hand. „Insbesondere für die Vokale, bei denen die Stimmlippen den Schall produzieren und der Vokaltrakt als akustisches Filter den Lautklang formt, ist die Analogie mit dem LPC-Sprachproduktionsmodell offensichtlich“[63].

5.3.3 Cepstrum

Weitere Merkmale liefern die Cepstralwerte. Um diese zu berechnen wird zuerst der Logarithmus des Spektrums gebildet. Danach wird die FFT noch einmal angewendet. Durch die doppelte FFT erhält man einen dem Zeitbereich parallelen Bereich. Dieser neue Bereich wird durch Vertauschen der Anfangsbuchstaben ausgedrückt: aus Spektrum wird Cepstrum, aus Frequenz wird Quefrenz, Filter wird zu Lifter.

Beim LPC-Verfahren lassen sich die Cepstral-Werte[64] c(k) direkt aus den Filterkoeffizienten berechnen. „Die cepstralen Koeffizienten mit den niedrigen Indizes (tiefe „Quefrenz”) (beschreiben)

[63] Beat Pfister, Tobias Kaufmann: Sprachverarbeitung – Grundlagen und Methoden der Sprachsynthese und Spracherkennung. Berlin 2017, S. 86 f.
[64] Der Unterschied zwischen LPC- und DFT-Cepstralwerten ist marginal.

die Grobstruktur des Spektrums, während die Koeffizienten mit höheren Indizes die Feinstruktur repräsentieren. Somit beschreiben die cepstralen Koeffizienten mit tiefen Indizes im Wesentlichen die Klangformung und damit die Charakteristik des Vokaltrakts, während die Koeffizienten mit höheren Indizes das Anregungssignal spezifizieren. Im cepstralen Bereich lassen sich also die Komponenten Signalproduktion und Klangformung, die gefaltet das Sprachsignal ergeben, näherungsweise wieder trennen."[65]

Eine Abwandlung dieses Verfahrens beschreibt das Mel-Cepstrum. Hierbei werden die Koeffizienten nicht auf Basis der DFT gebildet, sondern unter der Berücksichtigung psychoakustischer Experimente.

5.4 Akustische Bewertung

Die Hidden-Markov-Modelle (HMMs) sind die am weitesten verbreiteten Modelle zur Bewertung der extrahierten Merkmale. Dabei werden die Einheiten (siehe 5.2.3) durch in ihren Parametern angepasste Modelle repräsentiert. Für eine gegebene Zufallsfolge wird zunächst berechnet, mit welcher Wahrscheinlichkeit die unterschiedlichen Modelle diese Folge erzeugt haben könnten. Diese Wahrscheinlichkeitswerte werden später als Basis für die Erkennungsentscheidung genutzt.

Beobachtungen von HMMs können diskret oder kontinuierlich sein. Diskret bedeutet in diesem Zusammenhang, dass die Beobachtungswahrscheinlichkeit nur eine begrenzte Menge M annehmen kann. HMMs mit diskreten Beobachtungswahrscheinlichkeiten werden als DDHMMs[66] bezeichnet. Dem gegenüber stehen die kontinuierlichen Beobachtungen. Hier wird die Wahrscheinlichkeitsverteilung als Wahrscheinlichkeitsdichte beschrieben. Sie werden mit CDHMM[67] abgekürzt. Zur mathematischen Beschreibung von CDHMMs bietet sich die Gauss-Mischverteilung an.

Aufgrund der vorherigen Merkmalsextraktion sind die nötigen Merkmale bekannt. Allerdings sind die Laute selbst (und deren Wechsel) nicht mehr sichtbar. Es sind also nur die Beobachtungen und nicht die wahren Zustände bekannt. Daher kommt der Name Hidden-Markov-Modell. Es ist also möglich, die Wahrscheinlichkeitswerte endlicher Länge zu berechnen, ohne die Lautfolge zu kennen.

---

[65] Pfister, Kaufmann: Sprachverarbeitung, S. 92.
[66] engl. discrete density HMM.
[67] engl. continuous density HMM.

Wichtige Parameter für die HMMs λ sind die Zustandsübergangswahrscheinlichkeit a_{ij} und die Beobachtungswahrscheinlichkeit b_j. Damit gilt: λ = (A, B) mit A in der Darstellung einer N×N-Matrix der Zustandsübergangswahrscheinlichkeiten und B die Beobachtungswahrscheinlichkeitsverteilung in den Zuständen zwischen S_2 und S_{N-1}.

Abbildung 5-4 zeigt die schematische Darstellung eines HMMs. Dabei ist S_1 der Anfangszustand. D.h. in diesem Zustand wird die erste Beobachtung erst noch erzeugt. Der Zustand S_N stellt den Endzustand dar.

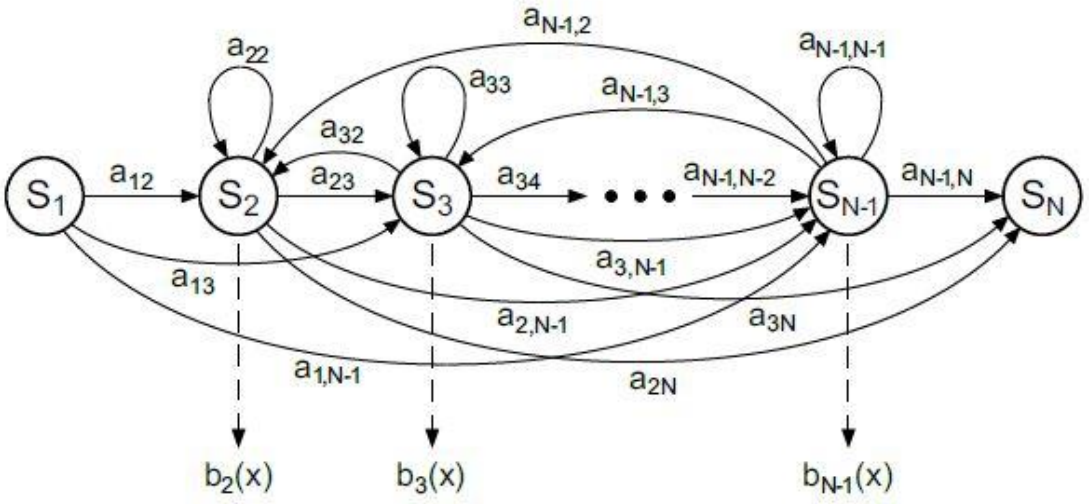


Abbildung 5-4: HMM 1. Ordnung mit N Zuständen.
(Quelle: Pfister, B., Kaufmann, T.: „Sprachverarbeitung – Grundlagen und Methoden der Sprachsynthese und Spracherkennung" (S. 110)).

5.4.1 Markov-Kette

Bei einer *Markov*[68]*-Kette* hängt die Wahrscheinlichkeit für einen Übergang nur vom vorhergehenden Zustand S_t ab und ist damit unabhängig von früheren Zuständen S_{t-1}, S_{t-2}, … S_{t-N}. oder einem Zeitpunkt t. Dies ist charakteristisch für die Markov-Kette erster Ordnung: Die Kette ist gedächtnisfrei. Typischerweise wird in der Spracherkennung eine „Rechts-Links"- Topologie verwendet. Dabei werden die Zustände von rechts nach links durchlaufen. Ein einmal verlassener Zustand kann nicht wieder betreten werden. Auch Überspringen ist nicht möglich.[69]

Dadurch gilt für die Übergangswahrscheinlichkeit von Zustand q_i nach q_j:

$$a_{i,j} = P(q_t = S_j | q_{t-1} = S_i), \qquad 1 \leq i, j \leq N.$$

[68] Benannt nach A. A. Markov (1856-1922), russischer Mathematiker.
[69] Überspringen ist in der Bakis-Topologie möglich. Diese wird ebenfalls in der Spracherkennung eingesetzt.

5.4.2 Probleme der HMMs

Für HMMs gibt es im Wesentlichen drei Probleme zu lösen:

1.) Berechnung der **Produktionswahrscheinlichkeit**: Hiermit wird die Wahrscheinlichkeit definiert, die eine gegebene Beobachtungssequenz P(X|λ) erzeugt. Dazu müssen die Wahrscheinlichkeiten P(X, Q|λ) mithilfe des Forward-Algorithmus über alle möglichen Zustandssequenzen *Q* aufsummiert werden.

2.) Die Berechnung der **Zustandssequenz** $\widehat{\boldsymbol{Q}}$, die eine gegebene Beobachtungssequenz *X* mit der größten Wahrscheinlichkeit erzeugt. Damit kann $\hat{Q}$ als optimale Zustandssequenz bezeichnet werden. Der Viterbi-Algorithmus ist dafür geeignet.

3.) Die **Parameter** des HMMs müssen so bestimmt werden, dass die Trainingsbeispiele möglichst gut beschrieben werden. Mithilfe des *Baum-Welch-Algorithmus*[70] (Forward-Backward-Algorithmus) können die Parameter so bestimmt werden, dass die Produktionswahrscheinlichkeit P(X|λ) dieser Beobachtungssequenz maximal wird.

Die Herleitung und Lösung der genannten HMM-Probleme muss nun getrennt für diskrete und kontinuierliche Signale berechnet werden. Allerdings führt dies für die vorliegende Arbeit zu weit.[71]

5.5 Dekodierung

Zur Definition der zu modellierenden Einheit stehen die Ganzwortmodelle stehen den Phonemmodellen gegenüber. Die Ganzwortmodelle kommen meist bei kleinen, spezialisierten Wortschätzen zum Einsatz, da hier für jedes Wort ein eigenes HMM gebildet wird. Dafür ist kein Wissen über die innere Struktur des Wortes notwendig. Allerdings ist eine Erweiterung des Wortschatzes mit erheblichem Aufwand verbunden. Außerdem wird die Tatsache, dass dieselben Laute in verschiedenen Worten benutzt werden, nicht ausgenutzt. Dadurch ergibt sich ein unnötig großer Speicherbedarf.

Im Vergleich dazu ist das Phonemmodell flexibler: Aus den Einheiten Phoneme und (Halb-) Silben können dem Baukastenprinzip folgend Wörter zusammengesetzt werden. Es hat sich bewährt, für häufig vorkommende kurze Wörter (beispielsweise Pronomen oder Artikel) auf das Ganzwortmodell zurückzugreifen und für alle anderen Wörter das Phonemmodell anzuwenden.

---

[70] Oder auch Viterbi-Training.

[71] Für weiterführende Beschreibungen empfiehlt sich: Pfister, Kaufmann: Sprachverarbeitung, S. 117 ff.

Schwierigkeiten ergeben sich aus den Abkürzungen (z. B. USA, ADAC, …), Zahlen (z. B. 1999) oder Ziffernfolgen (z. B. Agenda 2010). Hier muss im Vorhinein festgelegt werden, ob es sich um Wörter oder Buchstabenfolgen handelt. Problematisch sind ebenfalls Wörter mit mehreren Aussprachevarianten und eingedeutschte Wörter.

Für den Aufbau eines Wortlexikons können die Modelle der einzelnen Bausteine einfach aneinander gebaut werden (sofern es nur eine mögliche Transkription gibt). Damit entsteht ein großes HMM für ein Wort. Für die Übergänge zwischen den Zuständen (innerhalb der Wörter und auch zwischen den Wörtern) werden dazu nicht emittierende ε-Zustände eingeführt.[72] Nur dadurch kann im letzten Zustand die Wahrscheinlichkeit um den Betrag der Übergangswahrscheinlichkeit minimiert werden. Abbildung 5-5 zeigt den Anfang des Modells für das Wort „acht“.

Gibt es unterschiedliche Aussprachevarianten, kann jede Aussprachemöglichkeit in einem eigenen Modell gespeichert werden. Alternativ können bei nur geringfügiger Veränderung, die Varianten in einem Modell mit parallelen Zweigen realisiert werden.

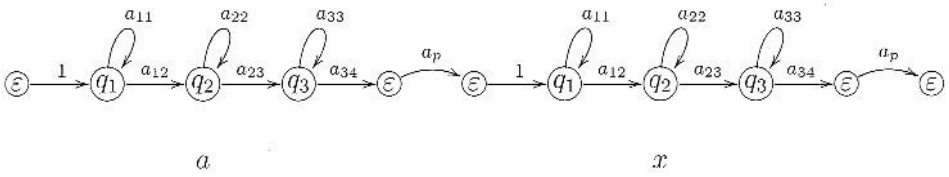


Abbildung 5-5: Anfang des Modells für das Wort "acht" mit ε-Zuständen.
(Quelle: Euler, S.: "Grundkurs Spracherkennung (S. 103)).

Ein statistisches Sprachmodell ordnet die möglichen Wörterkombinationen einer Wahrscheinlichkeitsverteilung zu. Zur Realisierung kann entweder auf syntaktische Regeln (Grammatiken) oder auf statistische Modelle (N-Gramm-Zeichen) zurückgegriffen werden.

5.5.1 Fehlerarten

Da in der Praxis kein System fehlerfrei arbeitet, werden Kenngrößen für die Fehlerwahrscheinlichkeit benötigt. Ein gründlicher Spracherkenner führt zu einem erfolgreicheren Gesamtergebnis. Damit ist die Erkennungssicherheit ein wichtiges Merkmal bei der Auswahl eines passenden Systems. Die Erkennungsgenauigkeit wird durch die Anzahl der Äußerungen und wie viele davon nach dem Test richtig erkannt wurden definiert. Aus der Anzahl der Fehler und der Gesamtanzahl der Wörter lässt sich der Erkennungsquotient errechnen.

[72] Diese Zustände besitzen kein Ausgabesymbol.

Um ein realistisches Ergebnis zu erhalten, ist es ratsam Verschiebungen zwischen Referenz und Ergebnis zu berücksichtigen. Deshalb werden folgende „Fehlerarten“ unterschieden:

- Verwechslung: Falsche Wörter wurden erkannt (N_V)
- Auslassung: an der Stelle gesprochener Wörter wird nichts erkannt (N_A)
- Einfügung: zusätzliche Wörter werden fälschlicherweise eingefügt (N_E).[73]

Darüber lassen sich Wortkorrektheit (WK) und Wortakkuratheit (WA) berechnen:

$$WK = \frac{N_W - N_V - N_A}{N_W} \quad \text{und} \quad WA = \frac{N_W - N_V - N_E - N_A}{N_W}$$

N_W: Gesamtanzahl der Wörter

Im Unterschied zur Wortkorrektheit berücksichtigt die Wortakkuratheit auch Worteinfügung. Der russische Mathematiker Vladimir I. Levenshtein veröffentlichte im Jahr 1965 ebenfalls ein Verfahren (Levenshtein-Distanz[74]), mit dem die minimale Anzahl der Fehler zwischen zwei Zeichenketten angegeben werden kann.

[73] Euler: Grundkurs Spracherkennung, S. 22.
[74] Vladimir I. Levenshtein: Binary codes capable of correcting deletions, insertions, and reversals. New York 1966, S. 845 ff.

6 Stand der Technik

Zu Beginn der Arbeit stand eine breitangelegte Recherche im Vordergrund, die sich nach und nach immer weiter auf die Themenbereiche Bild- und Audioschnitt fokussiert hat. Die interessantesten Ergebnisse dieser Recherche werden im Folgenden vorgestellt.

6.1 IBM Watson

Watson ist IBMs Supercomputer. Er ist auf die Beantwortung von Fragen in natürlicher Sprache spezialisiert. Namensgeber war Thomas J. Watson (Vorstandsvorsitzender von IBM in den Jahren 1949-1956). Dazu wird die natürliche Eingabesprache erfasst und in diversen Datenbanken nach Antworten gesucht. Da Watson ohne Internetverbindung arbeitet, werden alle strukturierten und unstrukturierten Daten auf einem Server gespeichert. Die Software-Engine ist „DeepQA“. Diese arbeitet auf dem Betriebssystem Linux.[75]

6.1.1 Jeopardy!

Um seine Intelligenz unter Beweis zu stellen, trat Watson im Februar 2011 in der TV-Show „Jeopardy!“ gegen menschliche Gegner an. Bei dieser Quizsendung werden dem Spieler Antworten aus verschiedenen Kategorien vorgestellt. Die Aufgabe des Spielers ist es nun so schnell wie möglich eine passende Frage zu formulieren. Die Anforderungen dieser Sendung sind damit gegensätzlich zu seinem sonstigen Anwendungsbereich. Watson löste diese Aufgaben in einem 3-Schritt-Modell:

a) **Linguistischer Präprozessor:** Analyse des Satzes und der Grammatik (via Parsing und Tokenizer) und Umwandlung in Prolog-Code

b) **Kandidatengenerierung**: Code wird an verschiedene Suchmaschinen weitergeleitet. So werden 100 bis 250 Suchergebnisse („Kandidaten“) als Antworthypothesen formuliert.

c) **Kandidatenbewertung**: Analyse der Hypothesen auf zeitliche und räumliche Zusammenhänge; ca. tausend Analysetools bearbeiten die Antwortmöglichkeiten, bevor ein Filter die Ergebnisse eliminiert, die keinen Beweis für die Korrektheit einer Antwort erbracht haben.[76]

---

[75] Vgl: Michael Loughran, Phil Zimmerman: „About IBM Watson – Factsheet“, unter: http://www-07.ibm.com/systems/hk/power/news/pdf/IBM_Watson_Fact_Sheet.pdf (Stand: 21.11.2017).

[76] Vgl. David Ferrucci, Eric Brown, u.a.: Building Watson: An Overview of the DeepQA Project. Palo Alto 2010.

Watson trat insgesamt dreimal gegen Sieger vorheriger Sendungen an. Im ersten Spiel waren Watson und sein menschlicher Kontrahent noch gleich auf. In den folgenden zwei Spielen allerdings gewann Watson mit großem Vorsprung.[77]

Die Teilnahme an „Jeopardy!" erinnerte an den Sieg von IBMs Schachcomputer „Deep Blue" gegen den Schachweltmeister Garri Kasparow am 11. Mai 1997 [78] . Dieser konnte das erste Match im Februar 1996 noch für sich entscheiden, verlor aber ein Jahr später in einer Revanche gegen DeepBlue. IBM rüstete den Computer extra dafür noch einmal deutlich auf. DeepBlue war so in der Lage, über 200 Millionen Stellungen pro Sekunde zu berechnen.

6.1.2 Trailer für Horrorfilm „Morgan"

Im Jahr 2016 wurde Watson beauftragt, den Trailer für den Science-Fiction-Horrorfilm „Morgan"[79] zu schneiden. Das Filmproduktionsunternehmen „20th Century Fox" war aufgrund der thematischen Nähe von künstlicher Intelligenz zur Handlung des Films an einer Kooperation interessiert.

Um Watson eine Vorstellung von Trailern aus dem Genre Horror-Film zu geben, wurde er mit 100 Horrorfilm-Trailern trainiert. Diese erforschte er, indem er zunächst jede Szene in Segmente unterteilte. Danach begann die eigentliche Analyse:

1.) Visuelle Analyse: Watson erkennt Personen, Objekte und Szenen. Jeder Person wird eine von 24 Emotionen und jeder Szene eine von 22 000 Kategorien (beispielsweise „ängstlich", „unheimlich") zugeteilt.

2.) Audio-Analyse: Untersuchung der Umgebungsgeräusche (Tonlage der Charaktere, Filmmusik,…), um jede Szene mit einem Gefühl beschreiben zu können.

3.) Analyse des Szenenaufbaus: Kameraposition, Einsatz von Lichteffekten werden betrachtet um diese Positionen als „typisch" für Horrorfilme zu kategorisieren.

Watson konnte nun erkennen, welche Szenen „üblicherweise" für Trailer dieses Genres verwendet werden. Danach wurde ihm das Filmmaterial von „Morgan" gegeben. Dieses Material analysierte er und wählte danach die zehn Sequenzen aus, aus denen der Trailer bestehen sollte. Watson benötigte dafür nur rund 24 Stunden, während Menschen für die Sichtung und der Auswahl des Materials

[77] Vgl. „O Watson no mundo", unter: http://www-03.ibm.com/marketing/br/watson/ (Stand: 21.11.2017).
[78] Weitere Informationen unter: http://www-03.ibm.com/ibm/history/ibm100/us/en/icons/deepblue/ (Stand: 21.11.2017).
[79] Dt. Filmtitel: „Das Morgan Projekt" (Regie: Luke Scott).

10 bis 30 Tage benötigen. Ein Schnittteam setzte danach den Trailer aus den ausgewählten Sequenzen zusammen.[80]

6.1.3 Cognitive Highlights

Im August/September 2017 wurde Watson Media zum ersten Mal beim Tennisturnier US Open eingesetzt. Das Tool „Cognitive Highlights" erkennt die wichtigsten Momente eines Spiels durch die Analyse der statistischen Daten, die Reaktionen der Zuschauer und Spieler, sowie der Gesichtserkennung.[81] Das System wählte anhand dieser Daten passende Szenen von den insgesamt sieben Spielfeldern aus und ordnete ihnen das Prädikat „Overall Excitement" zu. Im Schnitt wurden die Szenen zusammengesetzt. Damit blieb den Cuttern das mühsame und zeitraubende Sichten des Materials erspart.[82]

6.2 Smartphone-Technologie

Auch der Smartphone-Bereich hat die erweiterten Möglichkeiten, die sich durch den Einsatz künstlicher Intelligenz ergeben, erkannt. Die Einsatzgebiete erstrecken sich von Bildbearbeitung, Gesichtserkennung bis hin zur Akkulaufzeitoptimierung. Im Folgenden werden die neuesten Modelle von Huawei und Apple mit ihren Innovationen im Bereich der künstlichen Intelligenz vorgestellt.

6.2.1 Huawei Mate 10 Pro

Das Huawei Mate 10 Pro ist in Deutschland seit Mitte November 2017 auf dem Markt. Es enthält einen Kirin 970 KI-Prozessor mit einer Neural-network Processor Unit (kurz: NPU). Nach den Angaben von Huawei soll die Bilderkennung bis zu 25 Mal schneller arbeiten als die CPU und 50 Mal weniger Energie verbrauchen. Langfristig soll durch die NPU die Performance-Leistung gesteigert werden. Vorinstalliert ist der Microsoft Translator. Dieser übersetzt Textbausteine ohne Internetverbindung in über 40 Sprachen. Text auf Bildern wird ebenso erkannt und übersetzt. Bis jetzt funktioniert dies allerdings noch nicht mit gesprochener Sprache. Dies ist jedoch nur noch eine Frage der Zeit und könnte damit den Google Pixel Buds[83] Konkurrenz machen.

---

[80] Vgl. „IBM Research Takes Watson to Hollywood with the first 'Cognitive Movie Trailer'", unter: https://www.ibm.com/blogs/think/2016/08/cognitive-movie-trailer/ (Stand: 21.11.2017).

[81] Vgl. http://www-03.ibm.com/press/us/en/pressrelease/53068.wss#release (Stand: 21.11.2017).

[82] Vgl. Julio Fernandez: „Behind the Code: Cognitive Highlights at the US Open", unter: https://developer.ibm.com/dwblog/2017/cognitive-highlights-usopen/ (Stand: 21.11.2017).

[83] Kopfhörer, die via Bluetooth mit Google Assistent verbunden sind und mit einer Live-Übersetzungsfunktion ausgestattet sind.

Im Bereich der Bilderkennung wurde die Software mit 100 Millionen Bildern trainiert. Dadurch kann die Kamera in Echtzeit erkennen, was gerade im Sucher auftaucht. Der Bildausschnitt wird danach automatisch optimiert, um die Objekte bestmöglich abzubilden.[84]

Ein „intelligentes Akkumanagementsystem […] leitet […] eine personalisierte sowie adaptiv lernende Energieverteilungsstrategie ab, um die Langlebigkeit des Akkus zu maximieren."[85] Die Systemleistung des Smartphones wird so im Hintergrund immer weiter optimiert, um die Akkuleistung individuell zu maximieren.

6.2.2 iPhone X

Auch das iPhone X, das am 3. November 2017 in über 50 Ländern Verkaufsstart hatte, arbeitet mit künstlicher Intelligenz. Diese wird bei der Face-ID eingesetzt. Dieses Tool zur Gesichtserkennung wird zum Entsperren des Smartphones eingesetzt. Damit reicht ein kurzer Blick auf das Smartphone, um dieses zu entsichern. Mithilfe des A11-Bionic Prozessors verarbeitet die Neural Engine Informationen der TrueDepth-Kamera. Die Daten zur Gesichtserkennung werden zum einen aus einem Infrarotbild des Gesichts ermittelt und zum anderen wirft die True-Depth-Kamera 30.000 unsichtbare Punkte auf das Gesicht, um so ein 3D-Modell des Kopfes zu erstellen. Anhand dieser Merkmale, die im weiteren Verlauf in eine mathematische Darstellung umgewandelt werden, findet der Vergleich mit den registrierten Gesichtsdaten statt.

Das System soll auch Gesichter mit Bart, Mütze, Brille, usw. erkennen. Ebenso wird die Alterung der Benutzer berücksichtig. Face-ID funktioniert laut Apple sowohl in geschlossenen Räumen, als auch im Freien. Auch absolute Dunkelheit stellt kein Problem für die TrueDepth-Kamera dar.[86]

6.3 Adobe VoCo

Auf der Adobe Hausmesse MAX im November 2016 in San Diego wurde das Audiotool Adobe VoCo vorgestellt. Es handelt sich hierbei um eine Software, die Veränderungen an Sprachnachrichten vornehmen kann. Spiegel-Online Redakteur Jörg Breithut nannte das Tool „Photoshop für

[84] Vgl. Eric Ferrari-Herrmann: „Huawei Mate 10 Pro im Test: Macht nach einer Woche schon süchtig", unter: https://www.androidpit.de/huawei-mate-10-pro-test (Stand: 26.11.2017).
[85] „Huawei Mate 10 Pro", unter: http://consumer.huawei.com/de/phones/mate10-pro/ (Stand: 26.11.2017).
[86] Vgl. „Informationen zur fortschrittlichen Technologie von FaceID", unter: https://support.apple.com/de-de/HT208108 (Stand: 26.11.2017).

Audiodateien“.[87] Adobe VoCo entstand in Kooperation mit der Princeton University und soll Anfang 2018 entweder als Teil der Adobe Creative Suite oder als eigenständiges Produkt auf den Markt kommen.

Mithilfe des Programms lassen sich Sprachnachrichten sinngemäß verändern. Das bedeutet: die ursprüngliche Nachricht wird in einzelne Phoneme (siehe 5.2.3) unterteilt, die sich im nächsten Schritt in beliebiger Reihenfolge anordnen lassen. Damit lernt das Programm die Stimme eines Sprechers kennen, um sie später zu imitieren. Laut den Angaben der Entwickler wird dazu Audiomaterial mit einer Länge von 20 Minuten benötigt. Der Text kann danach so verändert werden, dass auch Wörter, die in der ursprünglichen Aussage gar nicht vorkamen, plötzlich gesprochen werden können. Dies wird durch eine einfache Eingabe des Textes in eine Eingabemaske realisiert.

In einem Youtube-Video[88] der Präsentation wurde die Originalaussage „Ich habe meine Hunde und meine Frau geküsst“ aus der Show „Key & Peele“ abgespielt und innerhalb weniger Sekunden in den Satz „Ich habe Jordan dreimal geküsst“ verwandelt.

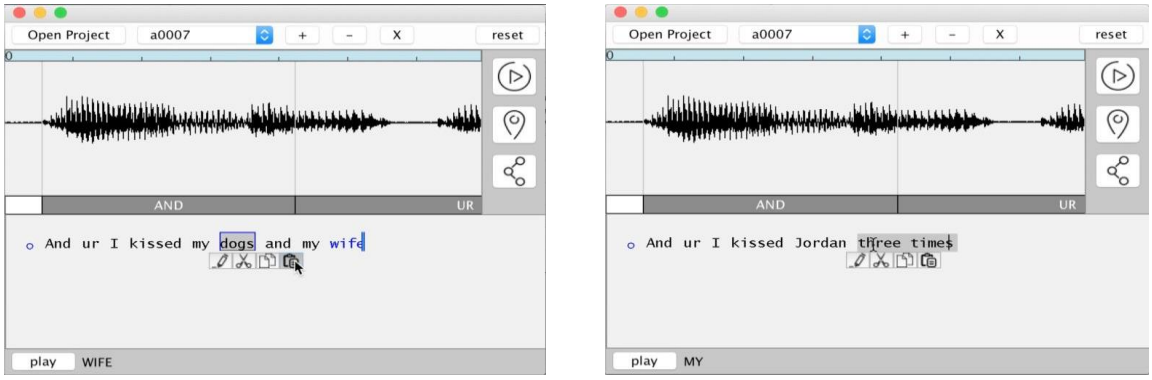


Abbildung 6-1: Funktionsweise Adobe VoCo (Quelle: https://techcrunch.com/2016/11/03/adobes-project-voco-lets-you-edit-speech-as-easily-as-text/ (Stand: 27.11.2017)).

Das Ergebnis ist absolut verblüffend: Das Original kann kaum von der manipulierten Aussage unterschieden werden. VoCo spielt den Satz flüssig und mit authentischer Stimme vor. Bei TechCrunch heißt es: „If you listen closely, you can hear when a word has changed, but it’s proba-

[87] Jörg Breithut: „Audio Editor: So stellt sich Adobe ein Photoshop für Sprache vor“, unter: http://www.spiegel.de/netzwelt/web/adobe-audio-das-photoshop-fuer-sprachdateien-a-1119783.html (Stand: 27.11.2017).

[88] Adobe Creative Cloud: „#VoCo. Adobe MAX2016 (Sneak Peaks) | Adobe Creative Cloud”, Youtube (04.11.2016), unter: https://www.youtube.com/watch?v=I3l4XLZ59iw (Stand: 27.11.2017).

bly only a matter of time before you won't be able to distinguish the actual recording and the edited (or completely fake) one."[89]

Damit könnten beispielsweise mühelos Versprecher und „Ähms" aus einer Audiodatei geschnitten werden, oder sogar ganze Hörbücher von einem Computer (aber mit jeder möglichen Stimme) eingesprochen werden.

Allerdings liegt genau darin das Risiko. Das Tool kann missbraucht werden, um Aussagen einen neuen Inhalt zu verpassen oder Bankgeheimnisse auszuspionieren, denn viele Banken und Versicherungen bieten den Kunden an, ihre Stimme als Passwort zu speichern.

Um diesen Problemen entgegen zu wirken, arbeiten Forscher bereits an „akustischen Wasserzeichen", die die Manipulationen am Audiomaterial aufdecken können. Denn selbst wenn das veränderte Audiomaterial nicht mehr vom Original zu unterscheiden ist, der Algorithmus hinterlässt immer Spuren.[90]

6.4 Google Clips

Google Clips ist eine Kamera, die selbst festlegt, wann sie auslöst. Clips bestimmt die besten Momente selbst und hält diese in kurzen Videoclips fest. Aus diesen kann der Nutzer später seine Bilder heraussuchen.

Mithilfe des maschinellen Lernens wird die Kamera während des Einsatzes immer besser. Sie optimiert anhand der Bildauswahl des Nutzers das Motiv und definiert

Abbildung 6-2: Google Clips-Kamera (Quelle: https://www.computerworld.com/article/3230132/mobile-wireless/google-s-clips-camera-offers-a-snapshot-of-things-to-come.html (Stand: 28.11.2017)).

[89] Frederic Lardinois: „Adobe's Project VoCo lets you edit speech as easily as text", unter: https://techcrunch.com/2016/11/03/adobes-project-voco-lets-you-edit-speech-as-easily-as-text/ (Stand: 27.11.2017).

[90] Vgl. Patrick Beuth: „Adobe: Tonaufnahmefälschen leicht gemacht", unter: http://www.zeit.de/digital/internet/2016-11/adobe-project-voco-photoshop-audio-manipulation (Stand: 26.11.2017).

so den „besten Moment“ nach den Wünschen des Nutzers um.[91] Nimmt die Kamera auf, leuchtet ein LED. Damit werden die Personen auf die laufende Aufnahme hingewiesen.

Da für die Nutzung keine Internetverbindung notwendig ist, verspricht Google Datensicherheit. Möglich ist der Einsatz der künstlichen Intelligenz durch den Einsatz des Movidius Myriad X-Chips. Dieser zeichnet sich durch eine hohe Rechenleistung aus. Die 12-Megapixelkamera ist mit einer 130°-Weitwinkellinse ausgestattet. Dadurch wird sichergestellt, dass ein großer Bereich abgedeckt wird.

Wann genau Google Clips auf den Markt kommt, wurde noch nicht bekanntgegeben. Bis jetzt ist nur bekannt, dass Google Clips ausschließlich in den USA vertrieben werden soll. Außerdem ist geplant, dass nur einige wenige Smartphones[92] Google Clips unterstützen werden.[93]

6.5 Audio Mining

Unter Audio Mining versteht man eine Technologie, die Audiosignale automatisch durchsucht und analysiert. Damit können einzelne Audiosignale in Redaktionsarchiven schnell und effektiv gefunden werden.

Das System des Fraunhofer-Instituts für intelligente Analyse- und Informationssysteme (IAIS) erkennt Schlüsselwörter und Phasen innerhalb des Audiosignals, die sofort in der Datenbank hinterlegt werden. Über eine Suchfunktion kann so direkt an die Stelle im Audiofile gesprungen werden, an der es um die gesuchte Thematik geht. Eine lange Suche durch eine Datenbank entfällt, da die Zeitinformation zu jedem Wort ebenfalls mitgespeichert wird. Gleichzeitig findet eine Analyse des Audiosignal statt. Dabei unterscheiden die Systeme nach den Kategorien Geräusche, Sprache und Musik. Es werden nicht nur Inhalte erkannt, ebenso findet eine Sprechererkennung statt.[94]

Die Spracherkennung wurde mithilfe des Toolkits „KALDI“ realisiert. Das Toolkit ist in der Apache-Lizenz frei verfügbar. Grundlage bilden Algorithmen des Deep Learning (siehe Kapitel

---

[91] Vgl. Martin Holland: „Google Clips: Smarte Kamera soll Fotografen aufs Bild holen“, unter: https://www.heise.de/newsticker/meldung/Google-Clips-Smarte-Kamera-soll-Fotografen-aufs-Bild-holen-3850038.html (Stand: 27.11.2017).

[92] Unterstützte Smartphones: Pixel-Smartphones, Samsung Galaxy S7 und S8, Apple iPhone 8 und 8 Plus.

[93] Vgl. Ingo Pakalski: „Google Clips: Mobile Kamera nutzt KI für Schnappschüsse daheim“, unter: https://www.golem.de/news/google-clips-mobile-kamera-nutzt-ki-fuer-schnappschuesse-daheim-1710-130439.html (Stand: 27.11.2017).

[94] Vgl. Audio Mining: Audiovisuelle Analysetools für Redaktionsarchive und die Medienbeobachtung, unter: https://www.iais.fraunhofer.de/de/geschaeftsfelder/content-technologies-and-services/uebersicht/AudioMining.html (Stand: 26.11.2017).

4.3). Durch die Verwendung von über 1000 Stunden transkribierten Materials arbeitet das System sprecherunabhängig und beherrscht Deutsch und Englisch mit insgesamt über 350.000 Wörtern. Wichtig in diesem Zusammenhang ist auch der Hinweis auf seine Echtzeitfähigkeit, dadurch kann das Audio Mining nicht nur im Archiv, sondern auch im Live-Betrieb angewendet werden.[95] Es ist also möglich die Analyse während einer Interviewaufnahme zu aktivieren. Die Ergebnisse (Transkript, Sprechererkennung) liegen unmittelbar nach dem Ende der Aufzeichnung vor. Der Schritt über das Archiv entfällt in diesem Fall.

6.6 Wibbitz

Die Internetplattform „Wibbitz" der israelischen Gründer Yotam Cohen und Zohar Dayan konvertiert Textnachrichten in Videoclips. Das Start-Up mit Sitz in Tel Aviv und Büros in New York und Paris arbeitete vier Jahre lang an der Technologie, bevor die Plattform 2013 eröffnet wurde.

Die Videos werden anhand von Nachrichten generiert. Das Tool wertet die Meldung aus und sucht speziell nach Personen und Sachverhalten. Um die Nachricht zu bebildern, durchforscht Wibbitz die großen Bildagenturen nach passendem Material und wandelt den Text in gesprochene Sprache um. Das Tool ist sogar in der Lage wichtige Informationen herauszufiltern und die Nachricht selbstständig zu kürzen.

Es muss nicht zwangsläufig das vorgegebene Bildmaterial verwendet werden, stattdessen kann auch selbst produziertes Material genutzt werden. Texteinblendungen oder Grafiken werden eigenständig erstellt. So werden insgesamt etwa 95% des Videos automatisch generiert. Das Feintuning bleibt dem Redakteur überlassen.[96]

Wibbitz-Videos bestehen also hauptsächlich aus einer Fotoshow, die mit einer Computerstimme angereichert wurde. Trotzdem ist das Start-Up erfolgreich, denn laut Gründer Zohar Dayan ist „es nicht sinnvoll, viel Zeit und viel Geld in Videos zu stecken. Nachrichten sind zu schnelllebig."[97]

---

[95] Vgl Dr. Joachim Köhler: „Suche und Navigation in Medienarchiven mittels Audio Mining", unter: https://www.bundesarchiv.de/imperia/md/content/bundesarchiv_de/fachinformation/informationstechnologie/digitalisiertesarchivgut/koehler_audiomining.pdf (Stand: 23.01.2018).

[96] Vgl. Ingo Georg: „Mit Wibbitz vom Text zum Nachrichtenvideo in 10 Sekunden", unter: http://www.osk.de/blog/mit-wibbitz-vom-text-zum-nachrichten-video-in-10-sekunden (Stand: 28.11.2017).

[97] Susanne Wittlich: „Der Bilder-Roboter", unter: http://www.focus.de/kultur/medien/kultur-und-medien-der-bilder-roboter_id_4745387.html (Stand: 28.11.2017).

7 Tools im Anwendungstest

In diesem Kapitel werden Tools und Software vorgestellt, die von der Thematik und dem Anwendungsgebiet her in einen Rundfunksender passen. Dabei geht es sowohl um die Bereiche Video- und Audioschnitt, als auch um Cloud-Computing und Roboterjournalismus.

7.1 Roboterjournalismus

Unter Roboterjournalismus versteht man Algorithmen, die aus Daten Texte produzieren können. Diese Textgenerierung hat den Vorteil, dass sich die Redaktionen auf die höherwertigen Texte und Berichte konzentrieren können. Währenddessen übernimmt der Mechanismus des maschinellen Lernens die Texte, die aus Zeitgründen sonst wegfallen oder eine zu kleine Zielgruppe ansprechen würden. Bis jetzt werden Roboterjournalisten häufig in den Bereichen der Sportberichterstattung, der Wettervorhersage und den Börsennachrichten eingesetzt.

Vereinfacht ausgedrückt werden aus vorher formulierten Bausteinen Texte gebildet, die mit den Daten vervollständigt werden. So können in Sekundenschnelle eine große Anzahl individueller Texte geschrieben werden. Damit ist beispielsweise das Schreiben der Wettervorhersage für jede einzelne Postleitzahl keine Kostenfrage mehr. Aufgaben wie diese werden Andreas Graefe zufolge von der Software schneller, billiger, und mit weniger Fehlern produziert, als mit einem menschlichen Journalisten.[98]

Gerechtfertigt werden kann dieser Ansatz mit den Ergebnissen einer Studie von Andreas Graefe und Mario Haim von der Ludwig-Maximilians-Universität (LMU) München. In dieser Studie wurden 681 Teilnehmer zu den Unterschieden zwischen einem maschinenerzeugten und einem von Menschen geschriebenen Text befragt. Abgesehen davon, dass die Leser – dank der neuesten Software – nicht klar zwischen dem Text eines Redakteurs und dem einer Maschine unterscheiden konnten, bevorzugten die „Participants […] human-written over automated news in terms of readability. For quality, the error bars crossed the mid-point of the scale, which means that differences were negligible. For credibility, participants favoured automated news. “[99] Als Grund dafür vermuten die beiden Forscher das häufigere Auftreten von Zahlen (mit Nachkommastellen) und Fakten.

[98] Andreas Graefe: Guide to Automated Journalism. New York 2016, S. 9.
[99] Mario Haim, Andreas Graefe: Automated news: Better than expected? London 2017, S. 1044-1059.

Führend auf dem Markt der automatisierten Texterstellung in deutscher Sprache ist die NLG-Cloud von AX Semantics. Dieses Tool wird im Folgenden ausführlich vorgestellt.

7.1.1 NLG-Cloud

Das Unternehmen AX Semantics (bis Oktober 2017 unter dem Namen aexea GmbH bekannt) wurde im Jahr 2001 von Saim Alkan in Stuttgart gegründet. Die Kommunikationsagentur hat von Beginn an Unternehmen im Bereich des Online-Content und der Online-Kommunikation beraten. Acht Jahre später begannen die ersten Versuche im Bereich der automatisierten Textproduktion. Aus dem Prototypen „Textautomat" wurde in den folgenden Jahren ein webbasierter Software-Dienst, der mittlerweile Texte in über 28 Sprachen übersetzen und Millionen von Texte pro Sekunde produzieren kann.[100] Die Software arbeitet zurzeit noch cloudbasiert. Allerdings gibt es Pläne für 2018, in denen sich die Unternehmen eine eigene Plattform konfigurieren können. Auch das Thema Datensicherheit hat bei AX Semantics einen hohen Stellenwert: „Wir laufen in der Amazon AWS und haben einen Vertrag mit Amazon gemacht […], der uns garantiert, dass wir in Frankfurt am Main […] (und) Rotterdam laufen."[101] Damit ist genau definiert, wo die Daten landen.

Die Natural Language Generation-Cloud (kurz NLG) ist ein Tool, das automatisch Texte mithilfe einer Maschine produziert. Grundlage hierfür bildet die eigens zu diesen Zweck entwickelte Auszeichnungssprache ATML3 (Automated Text Markup Language). Darin werden alle Daten, sowie deren logische Bedingungen gespeichert. Eine eigenständige Auszeichnungssprache ist deshalb so wertvoll, weil dadurch die Arbeit extrem erleichtert wird. „Wenn Sie versuchen alles in Python abzubilden, dann müssten so Dinge, wie die Morphologie einer Sprache, die grammatikalischen Zusammenhänge, die jetzt im Hintergrund realisiert werden […] jedes Mal neu implementiert werden", sagt Frank Feuler von AX Semantics[102].

Wie in Abbildung 7-1 dargestellt, werden für die Generierung eines Textes zunächst einmal strukturierte Daten benötigt. Im zweiten Schritt beginnt das Training. Hier werden die Texte eingepflegt und die logischen Verknüpfungen festgelegt. Danach werden die Daten und die Texte zusammengeführt und so der Text auf Basis der Daten produziert.

---

[100] Vgl. „Über AX Semantics", unter: https://www.ax-semantics.com/de/unternehmen.html (Stand: 02.12.2017).
[101] Frank Feulner: „Interview mit Frank Feulner (AX Semantics)". Anhang 3.
[102] ebd.

Abbildung 7-1: Bestandteile der Textgenerierung mit der NLG-Cloud. (Quelle: https://university.ax-semantics.com/library/ax-semantics-seminar/about/ (Stand: 04.12.2017)).

Die Daten müssen dabei verschiedenen Prüfungen standhalten. Sowohl im redaktionellen Bereich, als auch technisch, werden die Daten auf die Einhaltung der Regeln überprüft. Dies ist wichtig, um Schwierigkeiten im weiteren Verlauf aus dem Weg zu gehen. Besonders die Uneinheitlichkeit von Daten kann später zu erheblichen Problemen führen, die dann in der Masse untergehen könnten. Eine grobe Übersicht gibt Tabelle 7.

Redaktionell	Technisch
• Rechtschreibung • inhaltliche Verwertbarkeit • mögliche Ausprägungen • Befüllungsgrad	• Einheitlichkeit • Maschinenverarbeitung - Einträge in Grundform - durchgehend gleiche Einheiten - Listeneinträge mit einheitlichen Trennzeichen

Tabelle 7: Vgl. Datenanalyse (redaktionell/technisch).

In der Anwendung bedeutet dies, dass die Daten vorher genauestens auf Ungereimtheiten überprüft werden müssen. Das wird am folgenden Beispiel deutlich. Die untenstehenden Abbildungen geben Beispieldaten wieder.

ANALYSE: DATEN

Name	Hersteller	Kategorie	Display-größe (Zoll)	Gewicht (g)	Farbe	Betriebssystem	Prozessor-leistung (GHz)	Features	Prozessor-leistung Kerne
Samsung Galaxy Note 4	Samsung	Smartphone Samsung Galaxy Note 4	5,7	176	schwarz	Android	2,7	WLAN, Kamera und Bluetooth	vier
iPhone 7	Apple	Smartphone		150	Gold	iOS	2,7	Kamera	Quad-Core

Abbildung 7-2: Datenanalyse mit Fehlern (markierter Bereich).
(Quelle: https://university.ax-semantics.com/library/ax-semantics-seminar/about/ (Stand: 04.12.2017)).

Es wird deutlich, dass ein hohes Maß an Eindeutigkeit der Daten erforderlich ist um eine erfolgreiche Textgenerierung zu erreichen. Im obigen Beispiel (Abbildung 7-2) wurde unter Name sowohl der Produktname, als auch der Hersteller eingepflegt. Ähnliches ist in der Spalte Kategorie zu sehen. Hier wurde neben der Kategorie (Smartphone, Tablet, …) noch der Hersteller mit Produktnamen verwendet. In der Spalte „Farbe“ wurde „Gold“ versehentlich großgeschrieben.

Eine vollständige Korrektur der verwendeten Daten ist in Abbildung 7-3 zu finden.

ANALYSE: DATEN

Name	Hersteller	Kategorie	Display-größe (Zoll)	Gewicht (g)	Farbe	Betriebssystem	Prozessor-leistung (GHz)	Features	Prozessor-leistung Kerne
Galaxy Note 4	Samsung	Smartphone	5,7	176	schwarz	Android	2,7	WLAN| Kamera|Bluetooth	Quad-Core
iPhone 7	Apple	Smartphone		150	gold	iOS	2,7	Kamera	Quad-Core

Abbildung 7-3: Datenanalyse mit Korrektur (markierter Bereich).
(Quelle: https://university.ax-semantics.com/library/ax-semantics-seminar/about/ (Stand: 04.12.2017)).

Fehlende Daten können, müssen aber kein Problem darstellt. Werden ebenso Varianten der einzelnen Sätze programmiert, die das Fehlen eines Datenfeldes zulassen, wird der Text daraufhin aufgebaut.

Ausgehend von den Daten lässt sich der Aufbau einer automatisierten Textgenerierung grundsätzlich mit Abbildung 7-4 beschreiben:

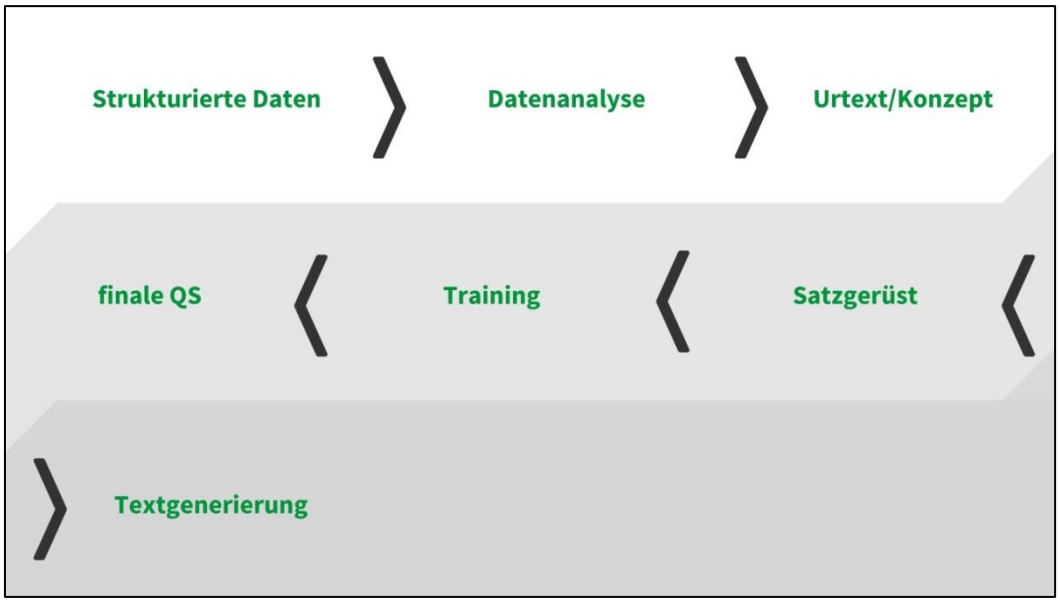


Abbildung 7-4: Von den Daten zum Text – die einzelnen Schritte.
(Quelle: https://university.ax-semantics.com/library/ax-semantics-seminar/about/ (Stand: 04.12.2017)).

Strukturierte Daten bilden die Grundlage für eine Datenanalyse nach vorher beschriebenen Regeln. Auf der Grundlage eines Datensatzes wird ein Urtext formuliert. Dieser beinhaltet einen Datensatz und dient als Ausgangsbasis für die weiteren Satzvarianten. Die Satzvarianten sind damit überwiegend paraphrasierte Aussagen des Urtextes.

Das Satzgerüst wird mit Varianten der Sätze und logischen Regeln gefüllt. Hier findet auch die Verknüpfung der Daten mit den Sätzen statt. Somit ist dieser Schritt für das nachfolgende Training von enormer Wichtigkeit. In den Sätzen werden sogenannte „Container" definiert. Diese Container stellen die Verbindung von Daten und Text her. Diejenigen Wörter eines Satzes, die durch andere Datensätze ausgetauscht werden können, werden zum Container. Das sind in der Regel alle Datenfelder. Dafür werden sogenannte „Properties" festgelegt. Diese helfen der Maschine zu entscheiden, welche Aussagen der Geschichte aufgrund der logischen Regeln wahr sind.

Die Properties werden durch drei Aussagen definiert:

- **truthExpression**: Treffen der logischen Entscheidungen und dementsprechende Freischaltung der einzelnen Textbestandteile. Sie muss immer einen Wert enthalten.
- **mappingExpression**: Auslesen der Daten aus dem Datensatz oder den Properties. Sie können hier allerdings auch verändert werden.
- **Vocabulary**: Festlegung der Ausgabe der Properties (Text im Satz).

Zum einfacheren Verständnis können die Properties als „Informationseimer“ angesehen werden. Daten werden über die mappingExpression („Input“) eingegeben. Als Schalter fungiert die truthExpression. In diesem Schritt findet die Wahrheitsüberprüfung statt. Je nach Ergebnis werden die Daten bei „true“ an den Text weitergegeben. Das Prinzip wird in der folgenden Abbildung 7-5 dargestellt. Gibt die truthExpression „false“ aus, wird nichts ausgegeben.

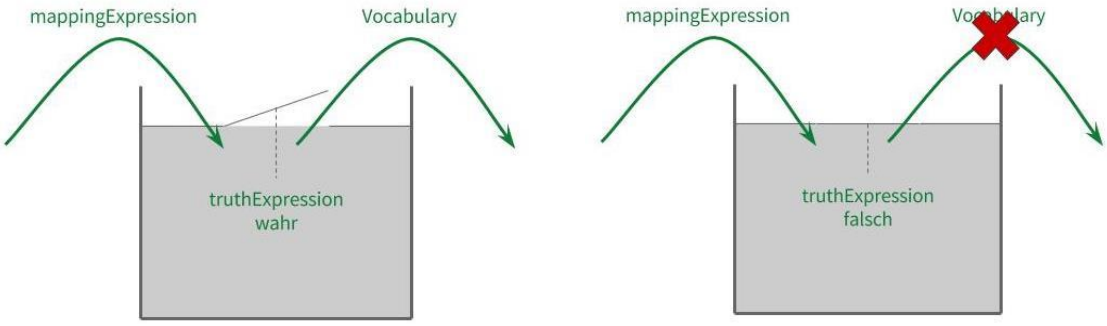


Abbildung 7-5: Funktionsweise der Property – true.
(Quelle: https://university.ax-semantics.com/library/ax-semantics-seminar/about/ (Stand: 04.12.2017)).

Im Anwendungsfall (Abbildung 7-6) bedeutet dies, dass das Datenfeld eines Datensatzes unter „groesse“ den Wert 6 enthält. Dieser wird über numeric(#groesse) eingelesen. Im folgenden Schritt wird überprüft, ob das Datenfeld nicht leer ist. Diese Aussage ist wahr. Damit wird der Wert weitergegeben und kann dann in einen Satz eingebaut werden.

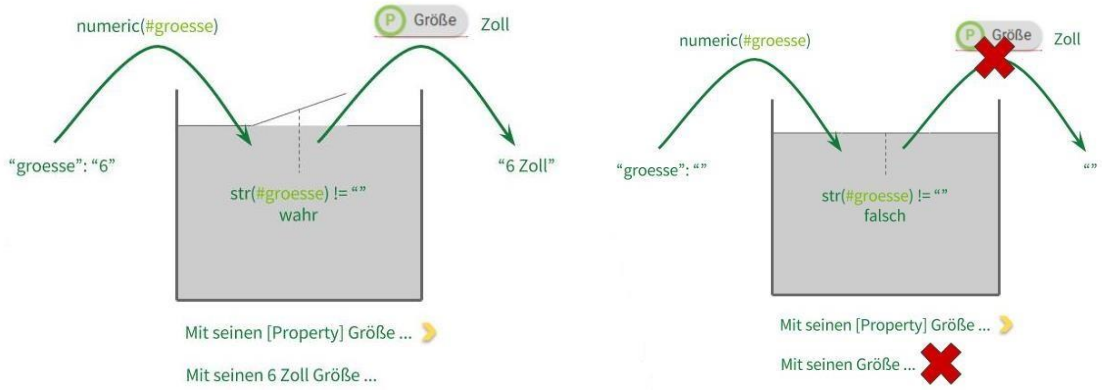


Abbildung 7-6: Funktionsweise der Property – false.
(Quelle: https://university.ax-semantics.com/library/ax-semantics-seminar/about/ (Stand: 04.12.2017)).

Demgegenüber steht der Fall (Abbildung 7-6), dass das Datenfeld tatsächlich leer ist. Daraufhin wird die truthExpression „false“ ausgeben und damit kein Wert weitergegeben. Dies bedeutet letztendlich, dass statt „Mit seinen 6 Zoll Größe …“ wie im ersten Fall, hier „Mit seinen Größe …“ ausgegeben wird.

Im Training werden die beiden Komponenten (Satz (-varianten) und die Logik) miteinander verbunden. Die potenzielle Ausgabe wird in der folgenden Qualitätssicherung auf Korrektheit überprüft. Daraufhin kann die Textgenerierung gestartet werden. Im Vergleich zur Textproduktion, muss die Generierung vieler Texte auf Grundlage eines Textes um einiges mehr leisten. Die weiteren Anforderungen werden in Tabelle 8 dargestellt.

Qualität eines Textes	**Qualität vieler Texte**
- angenehmer Textfluss - relevante Aussage - natürliche Sprache	- abwechslungsreich - Optimierung unterschiedlicher Eindrücke - einzigartig durch Datenbezug

Tabelle 8: Gegenüberstellung der Anforderungen an einen Text und an viele Texte.

Da natürlich nicht jeder Text gleich klingen soll und die Anmutung für die unterschiedlichen Zielgruppen berücksichtigt werden muss, bietet die Software unterschiedliche Möglichkeiten um die Varianz der Texte zu erhöhen. Dies kann zum einen durch das Tool „Lookup“ erreicht werden. Dies ist eine Art Nachschlagetabelle, die Strings (Zeichenketten) durch Synonyme ersetzen kann, um den Wortschatz anzupassen. Dadurch können sowohl „schönere“ Formulierungen verwendet werden als auch ein passender Wert in eine andere Sprache übersetzt werden.

Ein anderes Tool zur Erhöhung der Varianz ist der „Story Type“. Dabei kann die Reihenfolge der Sätze an den jeweiligen Kunden angepasst werden. Die Reihenfolge kann auch an bestimmte Bedingungen geknüpft werden. Dabei stehen die relevanten Informationen für den Benutzer im Vordergrund. Statt ihn mit Auskünften zu überhäufen, gilt es die interessanten Hinweise herauszufiltern und diese in ansprechender Weise zu präsentieren. Auch durch die „Properties“ kann ein Text abwechslungsreich gestaltet werden. Beispielsweise kann wie in Tabelle 9 dargestellt, je nach Definition der truthExpression und der mappingExpression die Speicherkapazität eines Smartphones unterschiedlich beschrieben werden:

< 16 GB	Mit einem Volumen von 8 GB ist der internen Speicher des Geräts vergleichsweise klein.
> 16 GB	Besonders punkten kann das Gerät in Sachen Speicher: Mit 64 GB Größe finden darauf sowohl Fotos, Videos und Apps als auch eine große Musiksammlung Platz.

Tabelle 9: Bedingungen bei Properties.
(Vgl.: Quelle: https://university.ax-semantics.com/library/ax-semantics-seminar/about/ (Stand: 04.12.2017)).

Damit ist die AX Semantics NLG-Cloud ein passendes Mittel um schnell eine große Anzahl an Texten zu produzieren. Auch individuelle Texte für eine kleine Zielgruppe können kostengünstig

geschrieben werden. Das Verhältnis zwischen der Reichweite und dem nötigen Aufwand wäre sonst nicht angemessen.

Besonders hervorzuheben ist die Benutzung der deutschen Sprache. Dies kann beinahe als Alleinstellungsmerkmal dieses Unternehmens angesehen werden. Andere Anbieter mit ähnlicher Software haben nur Englisch implementiert. Für einen deutschen Rundfunksender ist die deutsche Sprache allerdings ein wichtiges Kommunikationsmittel, auf das auch hier nicht verzichtet werden darf. Um den Algorithmus zu überwachen gibt es den ATML3-Debug-Tree. Damit kann für jeden generierten Text nachvollzogen werden, an welcher Stelle die Software welchen Wert berechnet hat.

Für die Zukunft ist die Nutzung neuronaler Netze in der Form geplant, dass die „Eingabesätze […] in Dependenzen zerlegt werden […] und alle grammatischen Zusammenhänge und Erbschaftsbeziehungen in diesen Sätzen annotiert, sodass sie nur noch Bezüge zu bestimmten Daten oder Vokabularen […] festlegen müssen.“ [103] Das neuronale Netzwerk analysiert also die natürliche Sprache und schlägt einen Regelsatz vor, der zur Generierung der Maschinensprache ATML3 aus der natürlichen Sprache nötig ist. „Es handelt sich dabei um eine hybride KI-Lösung, bei der ein Teil der Value-Chain durch echte, tief lernende und redaktionelle KI gelöst wird. Aber der Teil, der zur Laufzeit Texte produziert, die solide Automatisierung, das ist ein festes Regelwerk.“[104] Dieses Regelwerk soll mit der nächsten Generation der NLG-Cloud realisiert werden (voraussichtliche Veröffentlichung: März/April 2018). Die Sprachanalyse wird dann größtenteils von Google SyntaxNet auf Basis neuronaler Netze übernommen.

7.1.2 AX Semantics University

Unter AX Semantics University finden die Anwender passende E-Learning Kurse. Diese leiten den Benutzer durch die verschiedenen Schritte von der Erstellung eines Projektes bis zu dessen Fertigstellung. Neben Theorieteilen werden auch Praxisanwendungen gelernt. Damit kann der Nutzer die NLG-Cloud in seinem individuellen Tempo kennenlernen und ATML3 von Grund auf verstehen, um danach eigene Projekte anzulegen. Die Einarbeitungszeit bewegt sich nach den Erfahrungen von Frank Feulner im Bereich von wenigen Wochen.

Die E-Learning Kurse bieten eine gute Grundlage für einen ersten Einstieg in die NLG Cloud. Allerdings ist es teilweise schwierig, die einzelnen Schritte nachzuvollziehen. Dadurch ergeben sich

[103] Frank Feulner: Interview mit Frank Feulner (AX Semantics). Anhang 3.
[104] ebd.

Fragen, die teilweise nicht ohne eine lange Recherche gelöst werden können. Um diesem Problem entgegenzuwirken, ist die gesamte ATML3-Dokumentation[105] und die API-Dokumentation[106] im Internet zu finden. Des Weiteren gibt es ein Forum und ein Hilfezentrum, um sich mit Mitarbeitern oder anderen Schulungsteilnehmern auszutauschen.

Neben den E-Learning-Kursen bietet AX-Semantics auch mehrtägige Schulungen an. Damit will die NLG-Cloud ein Tool für Journalisten sein. Sie sollen auf der einen Seite bei ihren Fähigkeiten bleiben, „Geschichten zu erzählen, Informationen zu transportieren, mit der Audience zu interagieren und gewinnen einen neuen Teil dazu, der sich auf AX Semantics bezieht."[107] Dadurch gewinnt der Journalist einen Einblick in diverse IT-Konzepte und erlebt einen Business-Transfer zwischen den beiden Bereichen, die in dieser Software zusammenkommen.

7.2 Automatisierter Videoschnitt

Die Bilderkennung von Fotos ist mittlerweile auf einem guten Niveau. Mithilfe des Deep Learning (siehe Kapitel 4.3) können Katzen, Menschen, Gegenstände und auch Orte ohne Probleme auf Bildern erkannt werden. Dies funktioniert nach einem ähnlichen Prinzip wie Google Clips (Kapitel 6.4). Warum sollte dies dann nicht auch mit Videos funktionieren? Die heutige Computerleistung für 25 Einzelbilder pro Sekunde schafft das allemal. Auch die Gesichtserkennung und das Bewegungstracking stehen dem in nichts nach. Die Mevo-Kamera des US-amerikanischen Hersteller Livestream kann genau das.

7.2.1 Mevo-Kamera

Mevo ist eine 4K-fähige Kamera des New Yorker Unternehmens Livestream. Seit der Gründung im Jahr 2007 hat sich das Unternehmen zu einem der großen Streaminganbieter hochgearbeitet. Die Kamera kann in Verbindung mit einer App zum direkten Streaming auf die hauseigene Plattform von Livestream, sowie Facebook verwendet werden.[108]

Dabei hat die App nicht nur die Funktion des Streamings. Stattdessen erfolgt die gesamte Kamera- und Liveschnitt-Steuerung via App. Allerdings kann die Kamera auch autark arbeiten und das Video intern auf einer SD-Karte gespeichert werden.

---

[105] zu finden unter: http://docs.atml3.rocks/ (Stand: 10.12.2017).
[106] zu finden unter: https://developers.ax-semantics.com (Stand: 10.12.2017).
[107] Frank Feulner: Interview mit Frank Feulner (AX Semantics). Anhang 3.
[108] Mit Mevo Plus auch auf Youtube, Periscope und Twitter.

Das Bild (= Totale) der 4K-Kamera mit einem Sichtfeld von 150° kann in bis zu neun Bildausschnitte unterteilt werden. Diese Nahaufnahmen werden in 720p30 aufgezeichnet. Mithilfe der Gesichtserkennung und der Tracking-Funktion können Personen verfolgt werden. Dabei kann zwischen einem automatischen und einem manuellen Modus umgeschaltet werden.

Abbildung 7-7: Mevo-Kamera

(Quelle: https://www.adorama.com/images/Large/mv101ablb.jpg (Stand: 12.01.2018)).

Durch die Vorgabe einer Schnittgeschwindigkeit schneidet, schwenkt und zoomt die Kamera automatisch. Über den Touchscreen des Smartphones/Tablets können Zooms (Fingerspreizen), Schnitte (Tippen) und Schwenks (Wischen) selbst ausgeführt werden. Der Ton kann entweder durch das interne Stereo-Mikrofon oder über externe Mikrofone aufgezeichnet werden. Eine Aussteuerung wird intern realisiert. Auch Videofilter wie Sepia, schwarz/weiß, oder Belichtung und Kantenschärfe können eingestellt werden.

Unterhalb der Linse ist ein LED-Lichtring verbaut. Anhand verschiedener Farben können die Modi Aufnahme, Streaming, Akkuzustand und Netzwerk unterschieden werden. Die Verbindung zwischen Kamera und App wird wahlweise via WLAN oder Bluetooth realisiert.

7.3 Automatisierter Audioschnitt

Im Bereich des automatisierten Audioschnittes hängt vieles von der Spracherkennung (siehe Kapitel 5) ab, denn ohne die Umwandlung von Sprache in Text, macht der Audioschnitt in den seltensten Fällen Sinn.

Allerdings muss zwischen einem rein technischen Schnitt und einem inhaltlichen Schnitt unterschieden werden. Während es beim technischen Schneiden um die Entfernung der „ähms“ und Versprecher geht, wird beim inhaltlichen Schnitt je nach Aufgabenstellung die Aussage auf das Wesentliche reduziert. Beide Schnittarten sind selbstverständliche und alltägliche Arbeit in einem Hörfunkunternehmen. Der inhaltliche Schnitt wird von den Redaktionen übernommen und die

Aufgabe der Techniker ist die Entfernung von Versprechern und „ähms“. Vor allem der technische Schnitt kann – auch mit viel Übung und Erfahrung – extrem viel Zeit in Anspruch nehmen.

Ein Team um Dr. Aeneas Rooch des Unternehmens Scisys Deutschland GmbH hat sich dieser Herausforderung angenommen.

7.3.1 Scisys

Künstliche Intelligenz wird in vielen Bereichen eingesetzt, in denen entweder der Mensch ein zu hohes Sicherheitsrisiko mit sich bringt oder es sich um eine sehr monotone Arbeit handelt. Geht es stattdessen um Kreativität und Gestaltung, bleibt die Kontrolle in den meisten Fällen beim Menschen.

Genau solche Anwendungsfälle gibt es beim Audioschnitt. Die einfachen Fälle entsprechen hier der Situation, dass ein Text im Studio eingesprochen wurde und Anfang/Ende weggeschnitten werden müssen. Laut Aeneas Rooch ist das „das tägliche Geschäft eines Redakteurs und […] die Fälle, wo man wirklich diffizil schneidet, wo man wirklich ins Detail gehen muss, geschickt blenden muss, sind eher die seltenen Fälle.“[109]

Aber, um selbst einfache Schnitte der Software zu überlassen, muss diese selbst schneiden können. An erster Stelle muss eine Software dafür lernen, was einen guten Schnitt von einem schlechten Schnitt unterscheidet. Dafür werden die Regeln des maschinellen Lernens angewendet (siehe Kapitel 3). Das Ziel von Aeneas Rooch ist es zunächst einmal, eine Software zu entwerfen, bei der der Mensch einen Bereich angeben muss, in dem ein Schnitt gesetzt werden soll. Innerhalb dieses Bereichs sucht der Algorithmus die optimale Position aus. Dafür wertet der Algorithmus vorher aus, wo Menschen Schnitte setzen und lernt daraus, selbst eine gute Schnittposition zu finden.

Allerdings mussten dafür bereits im Vorfeld diverse Algorithmen auf ihre Plausibilität untersucht werden. Nicht jeder Algorithmus führt zum gewünschten Ergebnis. Selbst kurz bevor die Testdaten eines Senders verfügbar waren, waren sich die Wissenschaftlicher nicht sicher, ob es sich hierbei um einen passenden Algorithmus handelt. „Man trainiert (einen Algorithmus), indem man Datenbeispiele hinlegt und dazu verrät, wie man die einsortieren muss, ohne dass man verrät warum. […] Irgendwann, wenn man das anhand von einem großen Datensatz von Trainingsdaten gemacht hat, dann hofft man, dass man dem Algorithmus auch etwas vorsetzen kann und er macht die Klas-

[109] Aeneas Rooch: Interview mit Aeneas Rooch (Scisys). Anhang 5.

sifikation alleine anhand der Dinge, die er vorher gelernt hat."[110] Es handelt sich also um überwachtes Lernen (siehe Kapitel 3.1) mit dem Sonderfall, dass es hier nur Positivbeispiele gibt. Denn alle Schnitte werden als „gute Schnitte" bewertet, da die Wissenschaftler automatisch davon ausgehen müssen, dass die gesetzten Schnitte (der Menschen) gut sind.

Eine Herausforderung des Projekts lag in der Beschaffung der Daten. Denn sobald ein Audiofile geschnitten und abgespeichert wurde, sind alle Informationen über die gesetzten Schnitte verloren. „Man rendert ein fertiges Audio. Man führt alles zusammen und generiert ein neues Audio, in dem keine Informationen mehr über die alten Versatzstücke und Schnitte drin sind."[111]

Deshalb haben die Wissenschaftler eine Software entwickelt, die das Originalaudiofile mit den Informationen über die gesetzten Schnitte abspeichert. Dabei werden alle überflüssigen Daten verworfen um jeglichen Bedenken in Bezug auf Datensicherheit aus dem Weg zu gehen.

Die Qualität der Daten kann also gewährleistet werden. Allerdings ist trotzdem nicht klar, ob die gespeicherten Informationen überhaupt ausreichen. „Zum Beispiel können wir Mausbewegungen im Augenblick nicht nachverfolgen. Wir können also gar nicht sehen, wo jemand den Schnitt setzt und vielleicht nochmal umsetzt."[112]

Geht es nach Aeneas Rooch soll das Audiomaterial in Zukunft in einem Schnittfix ähnlich Microsoft Word geschnitten werden. In seiner Vision werden dabei zwei Technologien kombiniert: Die „Feinpositionierung der Schnitte" wie oben besprochen und die Spracherkennung (siehe Kapitel 5). Die Spracherkennung ist dafür verantwortlich, das Interview oder die Aufzeichnung zu transkribieren. Anhand dieses schriftlichen Materials können sich die Redakteure ihre O-Töne heraussuchen und zusammenbasteln. Durch die Verbindung von Audio zu Text ist zu jeder Sekunde klar, welche Textstelle welchem Zeitcode entspricht. „Dadurch, dass Sie die Wörter markiert haben, wissen Sie von wo bis wo das ungefähr in ihrer Audiodatei ist. Sie wissen den Start- und Endpunkt."[113] Diese Punkte werden dann von der Feinjustierungssoftware übernommen. „Dann bekommen Sie ein sehr sauber geschnittenes Audio, aus dem nur der ausgewählte Satz aus dem Transkript herausgeschnitten wurde. Das würde komplett überflüssig machen, dass wir mit der Hüllkurve arbeiten."[114] Aller-

[110] Aeneas Rooch: Interview mit Aeneas Rooch (Scisys). Anhang 5.
[111] ebd.
[112] ebd.
[113] ebd.
[114] ebd.

dings hängt alles vom Transkript ab. Beinhaltet dieses viele Fremdwörter, reden die Gesprächspartner durcheinander oder sind die Hintergrundgeräusche zu stark, wird die Erkennung nur eingeschränkt funktionieren. Handelt es sich aber um Studioaufnahmen, „brauchen Sie sich nicht mehr durch die Hüllkurve angeln. Sie können mit einem Blick den Text erfassen […] und können dann, wie in Word, einen Satz auswählen und diesen ausschneiden."[115]

Dafür ist Aeneas Rooch im Moment auf der Suche nach so viel Schnittmaterial wie möglich, denn nur so kann eine Software trainiert werden. Audiomaterial liegt den Rundfunkanstalten zwar zur Genüge vor, allerdings werden die Informationen über die Schnitte nicht gespeichert.

Mithilfe eines automatisierten Audioschnitts entstünde eine große Zeitersparnis sowohl auf Seiten der Redaktion als auch der Technik. Dadurch würden beide Abteilungen entlastet werden.

7.4 Cloud-Computing

Die großen weltweit agierenden Unternehmen haben Clouds entwickelt, mit denen die User sich ihre Applikationen auf Grundlage dieser Clouds selbst programmieren können. Tabelle 10 stellt eine Übersicht der fünf Unternehmen Amazon, Apple, Google, IBM und Microsoft mit deren jeweiligen KI und Cloud dar.

Unternehmen	**KI**	**Cloud**
Amazon	Alexa	Amazon Web Service
Apple	Siri	----------
Google	Google Assistant	Google Cloud Plattform
IBM	Watson	Bluemix
Microsoft	Cortana	Azure

Tabelle 10: Übersicht der Unternehmen, deren KI und Cloud.

Im Folgenden werden die Angebote der Bluemix-Cloud von IBM und der Azure-Cloud von Microsoft näher vorgestellt und deren Anwendbarkeit in einer Rundfunkanstalt kritisch betrachtet.

7.4.1 IBM: Bluemix

Die von IBM entwickelte Cloud nennt sich Bluemix. Die Entwickler können auf über 100 Anwendungen zugreifen und diese nach ihren Wünschen weiterentwickeln. Einen Auszug aus den angebotenen Cloud-Services im Bereich der KI ist in Anhang 6 zu finden.

[115] Aeneas Rooch: Interview mit Aeneas Rooch (Scisys). Anhang 5.

Besonders interessant für die vorliegende Arbeit sind die Tools „Visual Recognition“ und „Speech-to-Text“. Diese werden im Folgenden näher vorgestellt.

7.4.1.1 Speech-to-text

Mit dem der Autorin vorliegenden Tool „Speech-to-Text“ kann Sprache in Text umgewandelt werden. In der verwendeten Demoversion[116] (wie auch in der Vollversion) ist noch keine Transkription von Audiomaterial in Deutsch möglich. Stattdessen werden die Sprachen Portugiesisch (Brasilien), Französisch, Japanisch, Mandarin, Arabisch, Spanisch, Englisch (British und American) unterstützt. Das Tool kann über eine API mit der gewünschten Anwendung verbunden werden. Durch die Nutzung des maschinellen Lernens werden sowohl Informationen über eine Sprache und deren Grammatik, als auch über die Struktur eines Audiosignals verwendet. Eingespielt werden können Dateien in den Formaten Ogg oder Web Media (WebM), dem Opus oder Vorbis Codec, MP3 oder MPEG, Waveform Audio File Format (WAV), Free Lossless Audio Codec (FLAC) und Linear 16-bit Pulse-Code Modulation (PCM).

Die Software gibt neben dem Text auch einige zusätzliche Informationen wieder:

- **Speaker labels**: Erkennung von unterschiedlichen Sprechern
- **Word alternatives**: Wörter, die ähnlich klingen wie das transkribierte Wort
- **Word confidence**: Prozentsatz, der zu der Entscheidung für ein Wort geführt hat
- **Word timestamp**: Start und Ende eines jeden Wortes (Zeitmarke)
- **Keywords spotting**: Suche nach vorher festgelegten Schlagwörtern[117]

Um die Software zu testen, wurde für diese Arbeit unter einen kurzen Ausschnitt aus einem YouTube-Video diverse Geräusche und Musiken in unterschiedlicher Lautstärke gelegt. Zur Berechnung des Fehlerquotienten wurde auf die Berechnung unter Kapitel 5.5.1 zurückgegriffen.

Die folgende Untersuchung der IBM Watson Speech-to-Text – Software ist unter der Bedingung entstanden, dass die Software den Sprecher vorher nicht erlernt hat. Es handelt sich also um ein sprecherunabhängiges System.

Insgesamt wurden acht Testdurchläufe absolviert. Jeder Test hat eine neue Herausforderung in Form von verschiedenen Hintergrundgeräuschen an das System gestellt. Somit kann nachgewiesen

[116] Demoversion unter: https://speech-to-text-demo.ng.bluemix.net/ (Stand: 21.12.2017).
[117] Vgl. „IBM Cloud Documentation – Speech-to-Text”, unter: https://console.bluemix.net/docs/services/speech-to-text/index.html#about (Stand: 21.12.2017).

werden, bei welchen „Störsignalen“ das System die meisten Schwierigkeiten hat. Im Einzelnen bedeutet dies:

Test 1	Test 2	Test 3	Test 4	Test 5	Test 6	Test 7	Test 8
Sprecher	Atmo	Rauschen (leise)	Rauschen (laut)	instr. Musik (leise)	instr. Musik (laut)	Musik (leise)	Musik (laut)

Tabelle 11: Testdurchläufe bei Speech-to-Text.

Wie in Tabelle 11 dargestellt, beginnt der Test mit reinem Sprechersignal ohne Hintergrundatmo. Eine Straßenatmo wurde in Test 2 hinzugefügt. In den beiden folgenden Durchgängen wurde weißes Rauschen in zwei unterschiedlichen Pegelwerten dazu gemischt. In Test 5 und Test 6 wurde Instrumentalmusik und in Test 7 und 8 Musik mit Gesang jeweils in unterschiedlicher Lautstärke beigemischt. Damit sind die Situationen abgedeckt, die in einem Rundfunkunternehmen am häufigsten auftreten.

Ziel der Untersuchung ist die Berechnung der Wortkorrektheit und Wortakkuratheit.

Die Einstellung „Detect multiple Speakers“ wurde für die Untersuchung deaktiviert. Auf die Korrektur der Zeichensetzung wurde in der Fehleranalyse komplett verzichtet.

Um einen Einblick in den Originaltext und dessen Schwierigkeiten zu haben, ist dieser nachfolgend abgedruckt:

Originaltext:

Our world is changing in many ways and one of the things which is gonna have a huge impact on our future is artificial intelligence AI. Bringing another industrial revolution. Previous industrial revolutions expended humans mechanical power. This new revolution, this second machine age, is going to expand our cognitive abilities, our mental power. Computers are not just going to replace manual labor but also mental labor.[118]

Die Gesamtanzahl der Wörter (N_W) in diesem Text beträgt 67.

Im Folgenden werden nur die Ergebnisse von Test 1, Test 2 und Test 8 ausführlich besprochen. Die vollständigen Ergebnisse des Versuchs sind unter Anhang 7 zu finden.

[118] Benigo, Yoshua: „The Rise of Artificial Intelligence through Deep Learning”, unter: https://www.youtube.com/watch?v=uawLjkSI7Mo (Stand: 21.12.2017).

Test 1: Sprecher (Original)

Our world is changing in many ways and one of the things which is gonna have a huge impact on our future is artificial intelligence AI. Bringing another industrial revolution. Previous ~~industrialization~~ expended humans' mechanical power. This new revolution this second machine age. Is going to expand our cognitive abilities ~~are~~ mental power. Computers are not just going to replace manual labor but also mental labor.

Statt „industrial revolutions" hat das System „industrialization" erkannt. Das legt eine Vertauschung bei „industrial" und eine Auslassung bei „revolutions" nahe. Ebenso wird im weiteren Verlauf statt „our" „are" erkannt (Vertauschung). Daraus ergeben sich: $N_V = 2$ $N_A = 1$ $N_E = 0$.

Test 2: Sprecher mit Atmo

Our world is changing in many ways and one of the things which is gonna have a huge impact on our future is artificial intelligence AI. Bringing another industrial revolution. Previous doctrine revolutions expended humans mechanical power. This new revolution this second machine age. Is going to expand our ~~call me to. The teens are~~ mental power. Computers are not just going to replace manual labor but also ~~mentally~~.

Hier liegen Vertauschungen bei „doctrine", „call me", „are" und „mentally" vor. „to. The teens" werden als Einfügungen gewertet. Das letzte Wort „labor" wurde ausgelassen. Daraus ergeben sich folgende Verteilungen: $N_V = 5$ $N_A = 1$ $N_E = 3$.

Test 8: Sprecher mit Musik (laut)

~~Are we. The~~ way. One. ~~Just~~ gonna. ~~In fact. These~~ artificial. Another. ~~Daschle~~ revolution. ~~Expanded~~. ~~You~~.

Die Transkription fand hier sehr lückenhaft statt. Es wurden nur einzelne Wörter richtig erkannt. Oft wurden Wörter vertauscht bzw. ausgelassen. Naheliegende Einfügungen aufgrund des Gesangs in der Hintergrundmusik fanden kaum statt Das System kann offensichtlich gut zwischen den verschiedenen Stimmen unterscheiden. $N_V = 9$ $N_A = 52$ $N_E = 0$.

Schon anhand dieser drei Beispiele wird deutlich, dass die Ergebnisse sehr unterschiedlich ausfallen. Tabelle 12 liefert eine Übersicht aller Ergebnisse des Tests.

	Test 1	Test 2	Test 3	Test 4	Test 5	Test 6	Test 7	Test 8
	Sprecher	Atmo	Rauschen (leise)	Rauschen (laut)	instr. Musik (leise)	instr. Musik (laut)	Musik (leise)	Musik (laut)
N_V	2	5	4	17	6	7	8	9
N_A	1	1	0	6	1	34	1	52
N_E	0	3	0	4	2	1	3	0
N_W	67	67	67	67	67	67	67	67
WK	$\frac{64}{67}$	$\frac{61}{67}$	$\frac{63}{67}$	$\frac{44}{67}$	$\frac{60}{67}$	$\frac{26}{67}$	$\frac{58}{67}$	$\frac{6}{67}$
WK	95,5%	91,0%	94,0%	64,2%	89,6%	38,8%	86,6%	9,0%
WA	$\frac{64}{67}$	$\frac{58}{67}$	$\frac{63}{67}$	$\frac{40}{67}$	$\frac{58}{67}$	$\frac{25}{67}$	$\frac{55}{67}$	$\frac{6}{67}$
WA	95,5%	86,6%	94,0%	59,7%	86,6%	37,3%	82,1%	9,0%

Tabelle 12: Ergebnisübersicht IBM Speech-to-Text.

Insgesamt wird deutlich, dass das Ergebnis besser ist, je weniger Hintergrundgeräusche vorhanden sind. Dabei ist auffallend, dass leises weißes Rauschen bessere Werte liefert, als die Straßenatmo. Als komplett durchgefallen können die Tests 6 und 8 bewertet werden. Zu laute Hintergrundgeräusche haben zu großen Einfluss auf die Erkennungsrate, als dass mit diesen Ergebnissen gearbeitet werden könnte. Aus redaktioneller Sicht sind Ergebnisse im Bereich unter 80-90% nicht akzeptabel. Ansonsten bietet die Verwendung dieser Tools keine Zeitersparnis für die Redakteure.

7.4.1.2 Visual Recognition

Mit diesem Tool[119] können Bilder mittels Deep Learning-Tools analysiert werden. Daraufhin können Bilder in Bibliotheken abgelegt und kategorisiert werden. Visual Recognition von IBM Watson erkennt Gesichter, Landschaften und Objekte. Weiterhin kann es auf individuell festgelegte Erkennungsmuster trainiert werden.

Müssen Gesichter analysiert werden, macht die Software Angaben zum Alter und Geschlecht der Person auf dem Bild. Ebenso können Prominente erkannt werden. Datengrundlage ist ein IBM-

[119] Demoversion „Visual Recognition“, unter: https://visual-recognition-demo.ng.bluemix.net/ (Stand: 21.12.2017).

interner Server mit strukturierten und unstrukturierten Daten. Folglich liegt die Erkennungsquote ausschließlich an den Watson zur Verfügung gestellten Daten und nicht an dem freizugänglichen Internet.[120]

Um die Software zu testen, wurden jeweils drei Porträtbilder von zehn prominenten Schauspielern aus dem USA und Deutschland ausgewählt und diese der Software zugespielt. Bei der Zusammenstellung der Daten wurde auf eine Mischung verschiedener Altersgruppen und auf eine Verteilung von männlichen und weiblichen Schauspielern geachtet. Bei den Schauspielern handelt es sich im Einzelnen um:

- **USA**: Tom Cruise, Leonardo DiCaprio, Angelina Jolie, Heidi Klum, Brad Pitt
- **GER**: Veronika Ferres, Heike Makatsch, Til Schweiger, Matthias Schweighöfer, Nora Tschirner

In der Auswertung der Demoversion ist unter „ALTER" das vorgeschlagene Alter angegeben. Dieser wird in der Software noch mit einem Prozentsatz angegeben. Auf diesen wird in der Auswertung allerdings verzichtet, da kein Vergleich zu anderen Altersgruppen angegeben wird.

Unter „weiblich" und „männlich" wird dargestellt, zu welchem Anteil das Geschlecht erkannt wurde. Unter „Person" wird dokumentiert, ob die Person überhaupt erkannt wurde und mit welcher Sicherheit.

RUNDE 1:

	Cruise	**DiCaprio**	**Jolie**	**Klum**	**Pitt**
Alter	35 – 44	35 – 44	35 – 44	35 – 44	45 – 54
weiblich	----------	----------	1,00	0,99	----------
männlich	0,95	0,99	----------	----------	0,95
Person			1,00		0,95
	Ferres	**Makatsch**	**Schweiger**	**Schweighöfer**	**Tschirner**
Alter	35 – 44	35 – 44	35 – 44	0 – 17	18 – 24
weiblich	0,99	0,99	----------	----------	1,00
männlich	----------	----------	1,00	0,99	----------
Person		0,99	0,82	0,99	0,73

[120] Vgl. „IBM Cloud Documentation – Visual Recognition" unter: https://www.ibm.com/watson/developercloud/visual-recognition/api/v3/#create_a_classifier (Stand: 21.12.2017).

Es fällt auf, dass doppelt so viele deutsche Schauspieler richtig erkannt wurden, als Schauspieler aus den USA. Das Geschlecht wurde bei allen Personen richtig erkannt. Allerdings besteht beim Alter oft eine große Differenz zum wahren Alter der jeweiligen Personen.

RUNDE 2:

	Cruise	DiCaprio	Jolie	Klum	Pitt
Alter	35 – 44	35 – 44	35 – 44	35 – 44	
weiblich	----------	----------	1,00	0,99	----------
männlich	0,99	0,99	----------	----------	
Person	0,97	0,97	1,00	0,99	
	Ferres	**Makatsch**	**Schweiger**	**Schweighöfer**	**Tschirner**
Alter	35 – 44	18 – 24	35 – 44	0 – 17	18 – 24
weiblich	0,99	0,97	----------	----------	0,99
männlich	----------	----------	1,00	0,98	----------
Person			1,00		0,99

In der zweiten Runde hat sich nun ein völlig neues Bild ergeben. Auf dem Bild von Brad Pitt waren noch Wörter eines Werbebanners im Hintergrund zu erkennen. Dadurch war das System von den Wörtern so abgelenkt, dass überhaupt kein Gesicht erkannt wurde. Stattdessen erkannte das System Essen – was allerdings auch nicht korrekt ist.

RUNDE 3:

	Cruise	DiCaprio	Jolie	Klum	Pitt
Alter	35 – 44	25 – 34	35 – 44	35 – 44	45 – 54
weiblich	----------	----------	1,00	0,99	----------
männlich	0,97	0,95	----------	----------	0,95
Person					0,92
	Ferres	**Makatsch**	**Schweiger**	**Schweighöfer**	**Tschirner**
Alter	35 – 44	35 – 44	35 – 44	35 – 44	18 – 24
weiblich	0,99	0,99	----------	----------	0,99
männlich	----------	----------	1,00	0,99	----------
Person	0,95	0,99 + Roz Kelly		0,99	0,62

Die Auswertung der dritten Runde liefert ähnliche Personenerkennungsdaten wie in Runde eins: die Deutschen Schauspieler wurden besser erkannt. Allerdings ist die Erkennung grundsätzlich stark abhängig vom jeweiligen Bild. Beispielsweise wurden Tom Cruise und Leonardo DiCaprio jeweils nur in Runde 2 erkannt. Nora Tschirner hingegen in allen drei Runden.

Auffällig ist das Ergebnis der dritten Runde bei Heike Makatsch. Da sich in diesem Fall das System nicht sicher war und die Übereinstimmung für zwei Personen bei 99% lag, wurde zwei Personen für dieses Bild vorgeschlagen. Bei zu großer Unsicherheit ist die Entscheidung, um wen genau es sich handelt, also letztendlich dem Menschen überlassen.

Insgesamt liegt die Erkennungsquote über alle Durchläufe bei 17 von 30, was einem Wert von 56,67% entspricht. Die Erkennung des Geschlechts liegt bei 96,67%. Hingegen wird das Alter bei den Personen grundsätzlich eher zu jung eingeschätzt. Hier liegt die Erkennungsquote liegt bei 11 von 30. Dies entspricht einem Wert von 36,67%.

In Anhang 8 sind die Bilder in der jeweiligen Reihenfolge aufgelistet.

7.4.2 Microsoft: Azure

„Azure“ ist die Cloud-Computing-Plattform von Microsoft. Sie ist seit Februar 2010 online und bietet ihren Nutzern insgesamt über 150 Produkte an, die sich in 15 Anwendungsbereiche unterteilen lassen. Diese sind unter anderem: Datenbank, Daten & Analyse, Internet der Dinge, Netzwerk, Sicherheit & Identität, AI + Cognitive Services, sowie Überwachung & Verwaltung. Eine vollständige Liste der Anwendungen im Bereich „AI + Cognitive Services“ ist in Anhang 9 zu finden. Die Anwendungen im Bereich der künstlichen Intelligenz machen einen Anteil von etwa 20% des Gesamtangebots aus. Damit wird diesem Bereich ein hoher Stellenwert zugeordnet.[121]

Zur Vergleichbarkeit mit der Bluemix-Cloud von IBM werden auch hier das Spracherkennungstool „BING-Spracheingabe“ und die Bilderkennung „Maschinelles Sehen“ näher betrachtet und getestet.

7.4.2.1 BING-Spracheingabe

Die BING-Spracheingabe[122] übersetzt Sprache in Text und wieder zurück. Die Eingabe des Audiomaterials kann entweder über den Mikrofon-Input, über eine andere Audioquelle oder über eine

[121] Vgl. „Microsoft Azure Dokumentation“ unter: https://docs.microsoft.com/de-de/azure/#pivot=products&panel=all (Stand:31.12.2017).

[122] Demoversion unter: https://azure.microsoft.com/de-de/services/cognitive-services/speech/ (Stand: 31.12.2017).

Datei erfolgen. In allen drei Fällen analysiert das Tool die Audiodatei in Echtzeit. Dabei werden die Daten an einen Server gesendet und dann Teilergebnisse zurückgegeben. Die Dateneingabe erfolgt entweder über eine Programmierschnittstelle, die eine Netzwerkkommunikation zwischen Server und Client zulässt, eine sogenannte REST API, oder über eigens angelegte „Client libraries". Diese bereits konfigurierten Bibliotheken arbeiten unter Windows, Android und iOS und können in den Programmiersprachen C#, Java, JavaScript und ObjekctiveC implementiert werden.

Im Bereich der Spracherkennung werden drei unterschiedliche Modi angeboten:

- **Interactive Mode**: Der User stellt eine kurze Anfrage an das System, auf die die Software reagiert. Dem User ist zu jeder Zeit bewusst, dass er mit einer Maschine spricht.[123] Die Äußerung dauert üblicherweise nicht länger als 2-3 Sekunden und ist in einem gewissen Rahmen vorhersehbar. In diesem Fall kann mit Ganzwortmodellen gearbeitet werden (siehe Kapitel 5.5).
- **Conversation Mode**: Kontinuierliche Spracherkennung um die menschliche Kommunikation zu transkribieren. Dabei reagiert das System auf Alltagssprache mit längeren Sätzen sowie Sprachdialekte. Bei der Mensch-zu-Mensch-Kommunikation ist der Sprachwortschatz sehr groß, da nicht vorhersehbar ist, was gesagt wird. Für eine flexiblere Anwendung bietet sich hier das Phonemmodell an (siehe Kapitel 5.5).
- **Dictation Mode**: Dieser Modus ist zur Erkennung von längeren Äußerungen (Satzlänge 5-8s) gedacht. Dabei ist die Sprache eher formal und der User hat seine Äußerungen geplant. Die gesprochene Sprache wird dem User parallel als geschriebener Text angezeigt.

Die BING-Spracheingabe ist in der Lage 19 unterschiedliche Sprachen in interactive und dictation Mode zu verarbeiten (darunter Englisch, Deutsch, Spanisch, Französisch, Arabisch, Mandarin, Finnisch, Italienisch, Hindi und andere). Im conversation Mode können zehn Sprachen verarbeitet werden (Arabisch, Deutsch, Englisch, Französisch, Spanisch, Portugiesisch, Russisch, Italienisch, Japanisch und Mandarin).

Das Ausgabeformat kann ebenfalls vorher eingestellt werden. Zu unterscheiden sind die Formate „simple" und „detailed". Während im simple-Format lediglich der erkannte Text erscheint, gibt der detailed-Mode auch Informationen über die Alternativen aus. Denn nicht selten stellt sich die wahrscheinlichste Wortfolge im weiteren Verlauf als nicht sinnvoll heraus. In diesem Fall können

[123] Nicht zu verwechseln mit dem Turing-Test.

die Alternativen überprüft werden. Diese werden über eine N-best List[124] herausgearbeitet. Dabei wird eine Liste angefertigt, die Wörter nach deren Auftrittswahrscheinlichkeiten sortiert.[125]

Als besonderes Tool bietet die BING-Spracheingabe von Azure die Erkennung von Beschimpfungen und Beleidigungen in einer Aussage an. Für den Umgang mit diesen Ausdrücken gibt es drei verschiedene Möglichkeiten: 1.) Ausdruck durch Zeichen verbergen, 2.) Ausdruck löschen und 3) keine Änderung vornehmen.[126]

Um das Tool mit der Spracherkennung von IBM zu vergleichen, wurden dem System dieselben Audiofiles zugespielt. Die gesamten Ergebnisse sind in Anhang 10 zu finden.

Es wird wiederum nach der Fehleranalyse aus Kapitel 5.5.1 korrigiert, wobei Fehler in der Zeichensetzung nicht berücksichtig werden. An dieser Stelle werden nun die Ergebnisse des Test 1, des Test 2, und des Test 8 näher erläutert:

Test1: Sprecher

Our world is changing in many ways ~~than~~ one of the things which is gonna have a huge impact on our future is artificial intelligence, AI, bringing another industrial revolution. Previous industrial revolutions expended humans mechanical power. This new revolution this second machine age is going to expand our cognitive abilities ~~are~~ mental power. Computers are not just going to replace manual labor, but also mental labor.

In diesem Abschnitt wurden lediglich zwei Vertauschungsfehler erkannt: $N_V = 2$, $N_A = 0$, $N_E = 0$.

Test2: Sprecher mit Straßenatmo

Our world is changing in many ways ~~than~~ one. Artificial intelligence. Bringing another industrial revolution. Expended humans mechanical power. This is ~~Lucia~~ this second machine age is going to expand our cognitive abilities ~~are~~ mental power. ~~I'm~~ not just going to replace manual labor, but also mentally.

---

[124] Weitere Informationen zur N-best List unter: Yen-Lu Chow, Richard Schwartz: „The N-Best Algorithm: An Efficient Procedure for finding top N Sentence Hypotheses". In: Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP '90). Albuquerque 1990, S. 81–84.

[125] Vgl. Euler, Spracherkennung, S. 119.

[126] Vgl. „BING Speech Documentation" unter: https://docs.microsoft.com/de-de/azure/cognitive-services/speech/home (Stand: 30.12.2017).

Durch die Straßenatmo erhöhten sich die Vertauschungsfehler auf vier. Allerdings werden viele Wörter überhaupt nicht erkannt: $N_V = 4$ $N_A = 21$ $N_E = 0$.

Test 8: Sprecher mit Musik (laut)

~~*I will change it*~~. *Bringing another.* ~~*Expected. so*~~

In Test 8 wurden nur zwei Wörter richtig erkannt. Alle anderen Wörter wurden entweder vertauscht oder ausgelassen. Das letzte Wort „so“ wurde sogar eingefügt: $N_V = 5$, $N_A = 60$, $N_E = 1$.

Die Ergebnisse aller Tests sind in Tabelle 13 dargestellt.

	Test 1	**Test 2**	**Test 3**	**Test 4**	**Test 5**	**Test 6**	**Test 7**	**Test 8**
	Sprecher	Atmo	Rauschen (leise)	Rauschen (laut)	instr. Musik (leise)	instr. Musik (laut)	Musik (leise)	Musik (laut)
N_V	2	4	6	22	6	1	7	5
N_A	0	21	0	11	5	48	8	60
N_E	0	0	0	2	2	4	0	1
N_W	67	67	67	67	67	67	67	67
WK	$\frac{65}{67}$	$\frac{42}{67}$	$\frac{61}{67}$	$\frac{34}{67}$	$\frac{56}{67}$	$\frac{18}{67}$	$\frac{52}{67}$	$\frac{2}{67}$
WK	97,0%	62,7 %	91,0%	50,7%	83,6%	26,9%	77,6%	3,0%
WA	$\frac{65}{67}$	$\frac{42}{67}$	$\frac{61}{67}$	$\frac{32}{67}$	$\frac{54}{67}$	$\frac{14}{67}$	$\frac{52}{67}$	$\frac{1}{67}$
WA	97,0%	62,7%	91,0%	47,8%	80,6%	20,9%	77,6%	1,5%

Tabelle 13: Ergebnisübersicht Azure BING-Spracheingabe

Da die BING-Spracheingabe von der Azure-Cloud auch in der Lage ist Deutsch zu verstehen, wurde die Tauglichkeit in einem weiteren Test untersucht. Dazu wurde der Software den Beginn der Nachrichten von SWR2 (ohne Musikbett) und SWR3 (Schlagzeilen mit Musikbett) zugespielt.

Beispiel 1: SWR2-Nachrichten vom 01.01.2018 um 15 Uhr (Sprecher: Bert Linnemann)

Original:

15 Uhr: Verirrte Raketen oder der falsche Umgang mit Böllern haben in der vergangenen Nacht zahlreiche Brände ausgelöst. Mehrere Menschen kamen ums Leben oder wurden verletzt. In Heidesheim bei Mainz starb ein Mann bei einem Brand auf einem Campingplatz. Zwei weitere Männer kamen bei Unglücken mit Feuerwerkskörpern in Brandenburg ums Leben. Vor allem Kinder und Jugendliche verletzten sich beim Böllern. In Kaiserslautern wurde eine Frau von einem Auto erfasst und lebensgefährlich verletzt. Sie wollte auf dem Gehweg einen Feuerwerkskörper zünden.

BING-Spracheingabe:

~~Kippen~~ oder der falschen umgang mit böllern haben in der vergangenen Nacht zahlreiche Brände ausgelöst mehrere Menschen ~~kamenice~~ Leben oder wurden verletzt in heidesheim bei ~~meine~~ ~~Stadt~~ ein Mann bei einem Brand auf einem Campingplatz 2 weitere Männer kamen bei Unglück mit feuerwerkskörpern in Brandenburg ums Leben vor allem Kinder und Jugendliche ~~verletzt~~ sich beim böllern in Kaiserslautern wurde eine Frau von einem Auto erfasst lebensgefährlich verletzt sie wollte auf dem Gehweg einen Feuerwerkskörper ~~zu~~.

In diesem 80 Wörter umfassenden Beispiel sind ausschließlich Vertauschungen und Auslassungen zu erkennen. Insgesamt gilt: $N_V = 6$ $N_A = 6$

Dadurch ergibt sich eine Wortkorrektheit (WK) und Wortakkuratheit (WA) von 85%. Die Zeichensetzung und Rechtschreibung wurde nicht berücksichtig.

Beispiel 2: SWR3-Nachrichten vom 02.01.2018 um 11 Uhr (Sprecher: Christopher Jähnert)

Die getesteten Nachrichten von SWR 3 in Beispiel 2 wurden erst ab einem gewissen Abhörpegel erkannt. Der Grund liegt wahrscheinlich in dem Nachrichtenjingle, der in ein Musikbett übergeht. Über dieses Bett werden die Nachrichten von einem Nachrichtensprecher gesprochen.

Original:

(Jingle) Die Krise im Iran treibt den Ölpreis nach oben, Deutschland meldet einen neuen Jobrekord, ein Computerfehler macht Probleme am Frankfurter Flughafen. Das Wetter: windig mit einigen Schauern. 1 bis 9 Grad. (Bumper)

Die Proteste gegen die Regierung im Iran gehen weiter. Das Staatsfernsehen meldet, dass dabei seit gestern mindestens 9 Menschen ums Leben gekommen sind, hunderte sind schon festgenommen worden. Die Unruhen im Iran haben weltweit Auswirkungen auf den Ölpreis. Der steht jetzt so hoch wie zuletzt vor zweieinhalb Jahren berichtet SWR 3 Börsenkorrespondent Jan Plate.

BING-Spracheingabe:

Im Iran treibt den Ölpreis nach oben Deutschland meldet einen neuen Job Rekord ein computerfehler macht Probleme am Frankfurter Flughafen das Wetter windig mit einigen ~~schauen~~ ein bis 9 Grad. Die Proteste gegen die Regierung im Iran gehen weiter das Staatsfernsehen meldet das dabei seit gestern Mindestens 9 Menschen aus Leben gekommen Sind 100 E sind schon festgenommen worden die Unruhen im Iran haben weltweit Auswirkungen auf den Ölpreis der steht jetzt so hoch wie zuletzt vor 2 einhalb Jahren berichtet SWR 3 ~~Börsenkurs~~ ~~von NT~~ ~~an~~ ~~Platte~~

Das 85 Wörter umfassende Beispiel enthält Vertauschungen und gleich zu Beginn Auslassungen. Die korrekte Erkennung von Eigennamen wird nicht erwartet. Insgesamt gilt: $N_V = 3$ und $N_A = 2$. Die Wortkorrektheit liegt bei einem Wert von 94,1%. Damit ist dieser Wert deutlich höher als bei den reinen Nachrichten in Beispiel 1.

Auffällig ist, dass während der Erkennung bei der BING-Spracheingabe zunächst alle Wörter in Kleinbuchstaben geschrieben werden. Sobald das Audiofile beendet ist und damit die Spracherkennung stoppt, wird die Rechtschreibung allerdings nochmal angepasst. Die Satzzeichen werden in Echtzeit eingefügt. Damit liegt der Fokus während des Abspielens auf der reinen Spracherkennung und die Rechtschreibung wird erst im zweiten Schritt bearbeitet.

7.4.2.2 Maschinelles Sehen

Das Tool „Maschinelles Sehen“[127] ermöglicht es, Informationen aus Bildern zu filtern. Dadurch können die Bilder analysiert, kategorisiert und auch bewertet werden. Für die Analyse müssen die Bilder entweder hochgeladen werden oder via URL an den Server weitergeleitet werden. Dabei werden nur folgende Bildformate akzeptiert: JPEG, PNG, GIF und BMP. Das Bild darf eine Größe von 4 MB nicht übersteigen und muss mindestens die Maße 50 x 50 Pixel haben. Werden diese Vorgaben berücksichtigt, kann das Bild auf verschiedene Arten analysiert werden:

[127] Demoversion unter: https://azure.microsoft.com/de-de/services/cognitive-services/computer-vision/ (Stand: 31.12.2017).

- **Markieren des Bildinhalts**: Gegenstände, Lebewesen, Landschaft, Aktionen werden markiert und in einem beschreibenden Satze zusammengefasst. Die Markierungen beziehen sich nicht nur auf den Hauptgegenstand sondern auch auf das Geschehen im Hintergrund. Die Markierungen werden ohne eine Systematik vergeben.
- **Kategorisieren des Bildes**: Im Gegensatz zum schlichten Markieren des Bildinhalts wird hier das Bild anhand einer Systematik kategorisiert. Die Kriterien dafür sind in einer 86 Konzepte umfassenden Liste vermerkt.
- **Identifizieren des Bildtyps und seiner Qualität**: Handelt es sich bei dem Bild um ein schwarz/weiß- oder Farbbild, ist es eine Fotographie, eine Linienzeichnung oder ein Clipart?
- **Gesichtserkennung**: Erkennt Gesichter auf einem Bild und gibt die Koordinaten zurück, ebenso wie Geschlecht und Alter.
- **Spezialisierte Bildinformationen**: Erkennung von Prominenten und Sehenswürdigkeiten.
- **Bildbeschreibungen**: Zusammenfassung dessen, was auf dem Bild zu sehen ist. Die generierten Sätze werden auf Basis der berechneten Erkennungssicherheit erstellt.
- **Erkennung von nicht jugendfreien Bildern**: Anhand eines einstellbaren Filters können nicht erwünschte Bilder entfernt werden.
- **Schrifterkennung**: Befindet sich Text auf einem Bild, wird dieser erkannt und zur Weiterverarbeitung zur Verfügung gestellt. Die Sprache des Textes wird dabei automatisch erkannt. Unterstützt werden 25 Sprachen. Die Erkennungsgenauigkeit ist abhängig von der Bildqualität. Dies kann beispielsweise eine zu kleine Schriftgröße, dunkler Hintergrund, fehlende oder durchgestrichene Buchstaben/Wörter oder ein verschwommenes Bild sein. Handschrift kann grundsätzlich auch erkannt werden.
- **Erstellung von Thumbnails**: Um automatisch Vorschaubilder in unterschiedlichen Größen für die verschiedenen Plattformen zu generieren, bietet die Azure-Cloud einen eigenen Algorithmus an. Dieser analysiert das Bild nach dem Upload und sucht den interessantesten Bereich (region of interest) heraus. Im nächsten Schritt wird das Bild entsprechend zugeschnitten.[128]

Für einen vergleichenden Test zwischen der Bilderkennung von IBM und der Azure-Cloud, wurden dieselben Bilder auch dem Tool „Maschinelles Sehen“ von Azure zugespielt. Da die Azure-Cloud im Gegensatz zur Visual Recognition von IBM beim Alter ein genaues Jahr angibt (statt ei-

[128] Vgl. „Computer Vision Documentation“ unter: https://docs.microsoft.com/de-de/azure/cognitive-services/computer-vision/ (Stand: 31.12.2017).

nes 9-jahres Zeitraums,) wird darauf bei der Auswertung Rücksicht genommen: als korrekt werden alle Altersangaben bewertet, die im Bereich +5/-4 Jahre des geschätzten Alters der Person liegen.

Die Ergebnisse der drei Runden sind im Folgenden dokumentiert:

RUNDE 1:

	Cruise	DiCaprio	Jolie	Klum	Pitt
Alter	39	38	36	40	46
weiblich	----------	----------	0,95	o. A.	----------
männlich	0,99	0,99	----------	----------	0,99
Person	0,98	0,99	0,94	0,95	0,94
	Ferres	**Makatsch**	**Schweiger**	**Schweighöfer**	**Tschirner**
Alter	40	33	44	29	24
weiblich	o. A.	0,93	----------	----------	o. A.
männlich	----------	----------	0,94	0,93	----------
Person	0,98	0,98	0,94	0,98	0,94

Auffällig ist schon jetzt, dass im Vergleich zur Visual Recognition von IBM alle Prominenten erkannt wurden. Teilweise wurde allerdings bei der Sicherheit der Geschlechtsbestimmung kein Prozentsatz angegeben. Trotzdem wurde dieses richtig erkannt.

RUNDE 2:

	Cruise	DiCaprio	Jolie	Klum	Pitt
Alter	37	39	32	39	34
weiblich	----------	----------	0,92	o. A.	----------
männlich	o. A.	0,99	----------	----------	0,99
Person	0,91	0,96	0,95	0,93	
	Ferres	**Makatsch**	**Schweiger**	**Schweighöfer**	**Tschirner**
Alter	36	34	45	27	28
weiblich	0,96	o. A.	----------	----------	0,98
männlich	----------	----------	0,99	0,99	----------
Person	0,96	0,91	0,99	0,99	0,98

In Runde 2 wurden alle Schauspieler mit Ausnahme von Brad Pitt richtig erkannt. Auch die „Visual Recognition“ hatte bei dem Bild von Brad Pitt Probleme.

RUNDE 3:

	Cruise	DiCaprio	Jolie	Klum	Pitt
Alter	39	32	35	34	44
weiblich	----------	----------	o. A.	0,97	----------
männlich	0,99	0,99	----------	----------	o. A.
Person	0,98	0,97	0,97		0,99
	Ferres	**Makatsch**	**Schweiger**	**Schweighöfer**	**Tschirner**
Alter	40	33	35	43	29
weiblich	o. A.	o. A.	----------	----------	o. A.
männlich	----------	----------	0,99	0,98	----------
Person	0,98	0,98	0,50	0,96	0,92

Im Gegensatz zur „Visual Recognition“ von IBM hat „Maschinelles Sehen“ von Microsoft Heike Makatsch erkannt. Allerdings wurde Heidi Klum auf dem Bild lediglich als „a woman smiling for the camera“ beschrieben.

7.4.3 Auswertung

Im Folgenden werden die Ergebnisse der Spracherkennungstools und der Bilderkennung der beiden Unternehmen IBM und Microsoft näher betrachtet. Zunächst wird auf die beiden Spracherkennungstools Speech-to-Text (IBM) und BING-Spracheingabe (Microsoft) eingegangen, bevor es um den Vergleich der Tools Visual Recognition (IBM) und Maschinellen Sehen (Microsoft) geht.

Im direkten Vergleich liegt die Watson Speech-to-Text von IBM im Bereich der Spracherkennung deutlich vorne.

Lediglich bei dem reinen Sprachsignal weist die BING-Spracherkennung bessere Werte auf. Eine genauere Gegenüberstellung der Ergebnisse bietet die Übersicht in Abbildung 7-8.

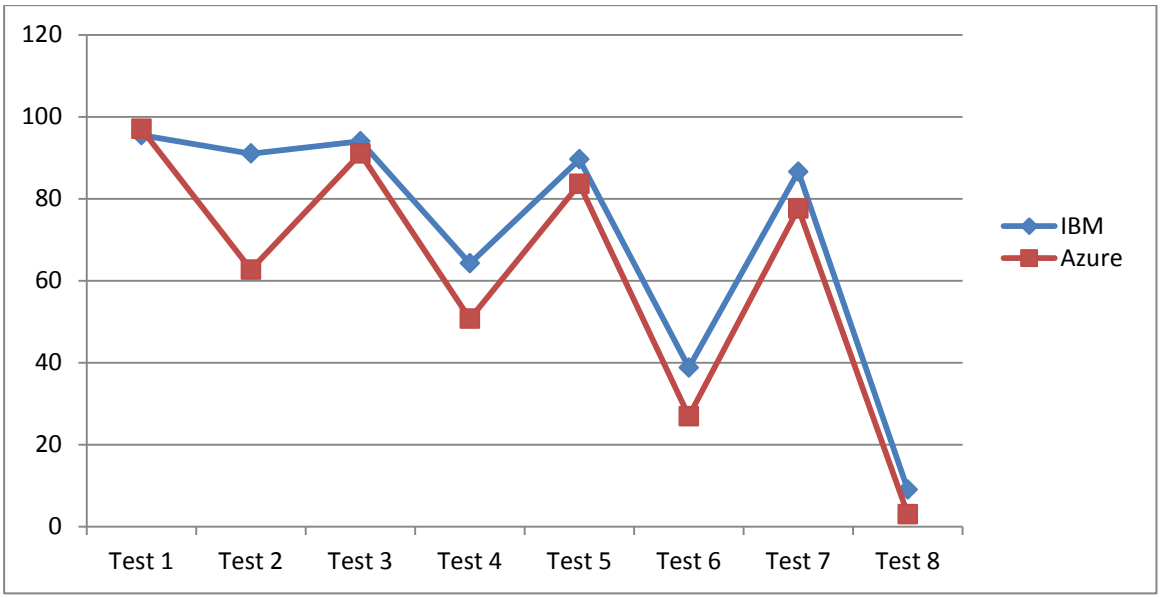


Abbildung 7-8: Vergleich Spracherkennung IBM <-> Azure.

Große Unterschiede sind lediglich in Test 2 zu erkennen. Offensichtlich hat Speech-to-Text weniger Probleme mit Hintergrundatmo, als die BING-Spracheingabe. In der Darstellung wurden nur die Ergebnisse der Wortakkuratheit (WA) dargestellt. Auch die Erkennung der deutschen Sprache liegt in den entsprechenden Bereichen. Damit sind Unterschiede in der Implementierung der verschiedenen Sprachen nicht zu erkennen.

Gerade im Hörfunk ist der Einsatz von Spracherkennungssoftware extrem hilfreich und zeitsparend, um Interviews oder Mitschnitte zu transkribieren. Hierbei sind aber nicht alle Interviews in einer Studioumgebung aufgezeichnet. Mit diesen Anforderungen muss die Software auf jeden Fall zurechtkommen, ansonsten kann sie sich nicht durchsetzen. Fraglich ist, ob ein Einsatz von sprecherabhängigen Systemen überhaupt sinnvoll ist. Sie würden zwar bessere Ergebnisse liefern, aber ob sich der Aufwand überhaupt lohnen würde, bleibt aufgrund der unüberschaubar großen Anzahl an Moderatoren und Redakteuren sowie einer noch größeren Menge an Interviewpartnern offen.

Die Erkennung von mehreren Sprechern und deren Verteilung ist ebenfalls ein interessantes und wichtiges Tool. Dieses ist bei der Azure-Cloud in einer anderen Anwendung[129] implementiert.

Bei der Bilderkennung überzeugt das Maschinelle Sehen von Microsoft. Es wurde lediglich Brad Pitt (Runde 1) und Heidi Klum (Runde 3) nicht erkannt. Generell fällt auf, dass die Prozentangaben im Vergleich zu den Angaben bei IBM sehr hoch sind. Das lässt auf eine hohe Sicherheit bei der Erkennung der jeweiligen Person schließen.

[129] Name der Anwendung: Sprechererkennung.

In der direkten Gegenüberstellung der Ergebnisse, ergibt sich Abbildung 7-9:

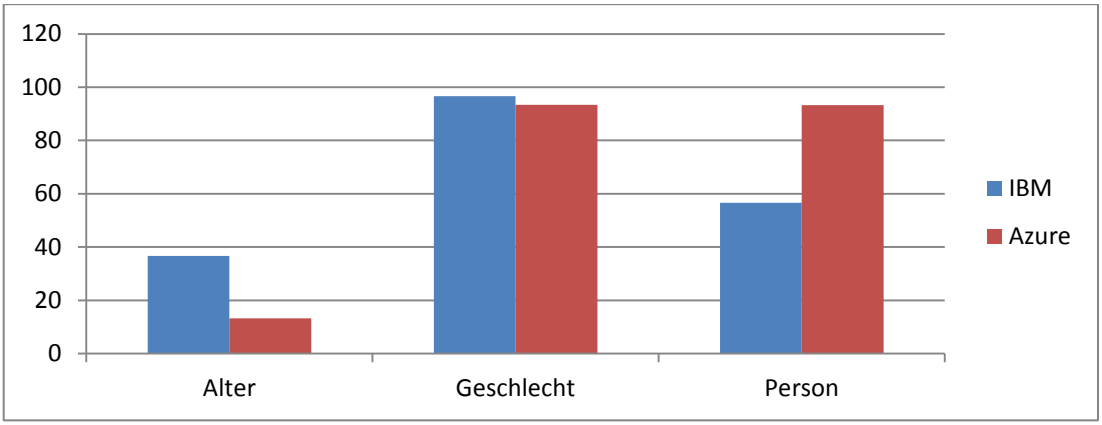


Abbildung 7-9: Vergleich Bilderkennung IBM <-> Azure.

Im Bereich der Erkennung von Personen im Allgemeinen und im Speziellen von Prominenten, überzeugt das Tool von Azure. Allerdings ist dies für einen Hörfunksender nur ein kleines Anwendungsfeld.

Wichtig für einen Hörfunksender ist die Bereitstellung von Metadaten. Damit können die Bilder zum einen in den Redaktionen schnell gefunden werden und an passender Stelle eingefügt werden. Zum anderen findet der Redakteur schneller eine bestimmte Information oder ein bestimmtes Bild, wenn ausreichend Metadaten vorhanden sind.

Um diese Metadaten zu genieren, ist eine umfassende Bilderkennung unumgänglich.

8 Anwendungsbereiche im Hörfunk am Beispiel des SWR

Die in Kapitel 6 und 7 vorgestellten und getesteten Tools werden im Folgenden auf ihre Anwendung im SWR erörtert. Dabei wird zunächst ein Statusbericht die aktuelle Situation beschreiben, bevor die Anwendungsmöglichkeiten in den einzelnen Bereichen vorgestellt werden.

8.1 Status Quo im SWR

Im Bereich des Hörfunks werden seit 2012 alle Datensätze in einer ARD-weiten Hörfunkdatenbank (HFDB) gespeichert. Diese Hörfunkdatenbank entstand in Kooperation zwischen ARD, Deutschlandradio und Deutsche Welle. Die Hörfunkdatenbank ihrerseits ist wiederum Teil der Normdatenbank (NDB). Federführend wurde die NDB vom Deutschen Rundfunkarchiv entwickelt und ist seit 2008 im Einsatz. Die Rundfunkanstalten einigten sich damit auf ein einheitliches System mit den Vorteilen, die ein zentraler Datenbestand mit sich bringt: Die Daten sind genormt und die Gefahr der Dopplung wird vermieden. Die Recherche-Oberfläche ist an die Arbeitsabläufe einer Redaktion angepasst und liefert durch die Schnittstellen zu anderen Datenbanken schnellen Zugriff auf jegliche Art von Datenmaterial. Durch die Zentralisierung der Datenbanken konnten zudem die laufenden Kosten gesenkt werden, da Entwicklung und Instandhaltung zentral bearbeitet werden.[130]

Somit können Informationen über Künstler und Institutionen, aber auch Wort- und Musikbeiträge von unterschiedlichen Quellen gesammelt und den anderen Rundfunkanstalten zugänglich gemacht werden.[131] Hierzu werden auch Altbestände wie Tonbänder, Schallplatten, DAT-Kassetten und CDs digitalisiert und kategorisiert. Das Pendant der HFDB im Bereich des Fernsehens ist die Fernseharchivdatenbank (FESAD).[132]

8.2 Archiv

Zur Archivierung der Daten (u.a. Wort- und Musikbeiträge) werden diese in den Audio-Massenspeicher (AMS) überspielt, der diese wiederum der HFDB zufügt. Da bis jetzt für jeden Datensatz eine große Anzahl an Metadaten eingegeben werden muss, besteht hier großer Hand-

---

[130] Michael Ranft: „Entwicklung einer Hörfunk-Archivanwendung“, unter: https://www.ordix.de/images/infomaterial/Referenzflyer/ordix_referenz_swr.pdf (Stand: 05.01.2018).

[131] Vgl. „Normdatenbank der ARD“, unter: http://www.ard.de/home/die-ard/fakten/abc-der-ard/Normdatenbank_der_ARD/486390/index.html (Stand: 05.01.2018).

[132] Vgl. Andreas Dan: „ARD Normdatenbank – Nutzung und Pflege von Normdaten in der ARD“, unter: http://in2n.de/medien/2014/08/20140930_IN2N-Workshop_Dan_ARDNormdatenbank.pdf (Stand: 05.01.2018).

lungsbedarf. Neben der Tatsache, dass die Metadaten auch von einer Software zuverlässig generiert werden können, ist auch die Verschriftlichung (Transkription) bzw. Verschlagwortung der Audiofiles von enormer Wichtigkeit, denn nur so können bestimmte O-Töne schnell und ohne langes Suchen gefunden werden. Mithilfe der Spracherkennung könnte direkt ein schriftlicher Text des Gesprochenen angelegt werden. Dieser wird bei der Suche direkt angesprochen und liefert so die passende Stelle in einem Audio-File. Der ganze Vorgang entspricht ungefähr dem Prinzip des Audio Mining (Kapitel 6.5).

8.3 Redaktion

Für die alltägliche Arbeit in einer Redaktion, besonders wenn es sich um eine tagesaktuelle Redaktion handelt, werden oft schnell O-Töne gesucht. Sind die vorhandenen Redaktionssysteme mit einer Audio Mining-Software verbunden, kann beispielsweise direkt an die passende Stelle gesprungen und der O-Ton zugeschnitten werden, ohne dass das komplette Interview angehört werde muss. Damit fällt das zeitraubende „Durchhören“ diverser Audiofiles weg. Stattdessen erkennt das System selbstständig die passenden Stellen. Wichtig ist allerdings, dass trotzdem das ganze File geladen und der Schnitt in diesem Fall nicht dem System überlassen wird. Denn sonst kann es schnell zu sinnverändernden O-Tönen kommen, da der Kontext eventuell nicht berücksichtig wurde. Damit funktioniert das Audio Mining ähnlich der Idee von Aeneas Rooch (siehe Kapitel 7.3.1), mit dem Unterschied, dass Aeneas Rooch auch den Schnitt dem System überlassen möchte.

Einen weiteren Anwendungsfall in einer Redaktion stellt der Roboterjournalismus dar. Dadurch können Texte in einer Menge produziert werden, die sonst so viel Zeit in Anspruch nehmen würde, dass das Verhältnis zwischen Kosten und Nutzen nicht mehr stimmt. Gerade für einen Sender mit regionalem Bezug kann sich dadurch eine neue Zielgruppe eröffnen, beispielsweise durch die Veröffentlichung der Sportergebnisse lokaler Vereine, lokaler Events und Veranstaltungen oder Verkehrsmeldungen.

8.4 Sendestudio

Im Bereich des Sendestudios ist eine Sprach- bzw. Sprechererkennung in mehreren Anwendungsfällen sinnvoll. So kann im Selbstfahrbetrieb eine Sprechererkennung automatisch das eingestellte Sprachsetup laden und EQ und Kompressor selbstständig an die Stimme anpassen um ein optimales Ergebnis zu erzielen.

Des Weiteren kann die Stichworterkennung helfen, vorproduzierte Sendungen ohne Techniker abzufahren. Bei SWR2 beginnt nahezu jede volle Stunde mit Nachrichten. Diese Enden mit der Ansage der Uhrzeit, beispielsweise „Fünfzehn Uhr fünf“. Mit einem Trigger auf „Uhr“, der mit ein bis zwei Sekunden Versatz das Abspielen der nächsten Sendung auslöst, startet die Sendung, ohne dass ein Techniker anwesend sein müsste. Ein ähnliches Prinzip kann auch bei dem Nachrichtensender SWR Aktuell angewendet werden. Dieser sendet Montag bis Freitag zwischen 6 und 18 Uhr live im Selbstfahrbetrieb, in der Nacht und am Wochenende wird das Programm ausschließlich aus der Automation gestartet. Gerade während dieser Automationszeiten können so Sendelöcher vermieden werden.

Im Verlauf dieser Arbeit hat sich ein weiteres Anwendungsgebiet aufgetan: die Vermeidung eines Sendeausfalls. Wie Abbildung 8-1 zeigt, kommen im Mischpult die Signale des Mikrofons, des RadioMax (Musik) und des NewsMax (Beiträge) zusammen. Die Verknüpfung verschiedener Systeme birgt ein gewisses Risiko. Für die Nacht oder am Wochenende wird der RadioMax programmiert. Nur so kann ein reibungsloser Übergang zwischen den einzelnen Sendestunden gewährleistet werden. Dies ist vor allem während der Nachtautomation wichtig. Allerdings ist an dieser Stelle die Wahrscheinlichkeit für das Auftreten eines Fehlers sehr hoch. Die Programmsumme wird im weiteren Verlauf an den Schaltraum weitergeleitet. Von dort aus wird das Signal an die Sendestation weitergeleitet.

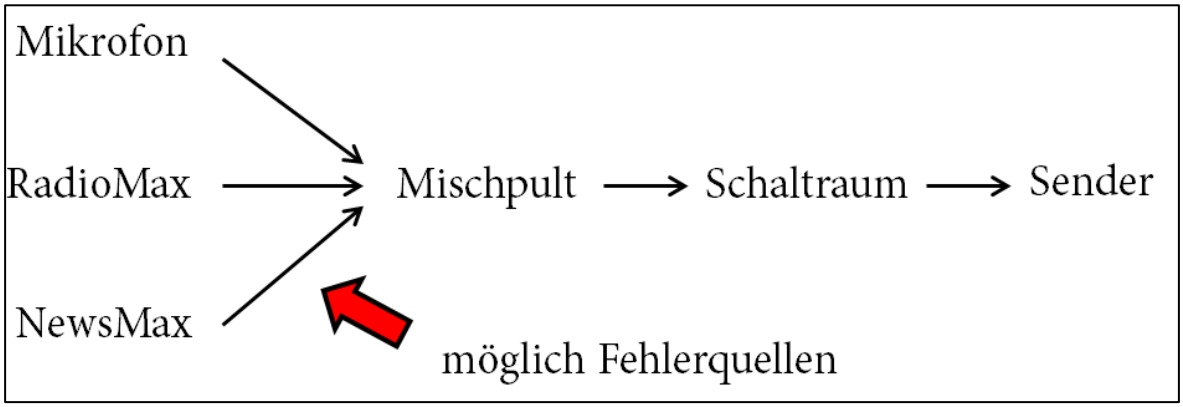


Abbildung 8-1: Signalweg

Der Schaltmeister bekommt allerdings erst nach 20 Sekunden „Stille“ eine Mitteilung über einen Sendeausfall. Er reagiert also erst dann, wenn der Sendeausfall schon mindestens 20 Sekunden andauert. Deshalb kam im Verlauf meiner Arbeit die Idee auf, ein System zu entwickeln, das schon vor dem eigentlichen Ausfall Probleme und Fehler erkennt. Dafür wurden von Dominik Mayer im

Windows AutoHotKey die bereits gemachten Fehler hinterlegt. Nach dem Anlegen der Sendestunden für die unmoderierten Zeiten (Nacht und Wochenende) prüft die Software, ob die Planung Fehler enthält. Durch dieses System ist es nun möglich, den Fehlerquotienten immer weiter zu minimieren, um Sendungsausfälle zu verhindern.

8.5 Visual Radio

Unter dem Begriff „Visual Radio“ verkauft das Unternehmen Chyronhego eine Software, die Kameras in Sendestudios oder Gesprächsrunden steuern kann. Dabei können drei unterschiedliche Steuerungsmodi unterschieden werden: manuell, halb-automatisch oder voll-automatisch. Das Visual Radio wird derzeit bei den Sendern SWR3 und DASDING eingesetzt.

Über die Software „Audio Director“ können die Remote-Kameras bedient werden, Bildausschnitte festgelegt und Grafiken (beispielsweise Bauchbinden) oder Animationen (Logo, Trenner) eingepflegt werden. Neben einem Live-Stream aus dem Studio können Tweets, Facebook- oder Instagram-Posts eingebunden werden. Auch Informationen zum aktuellen Song oder den Künstlern (über Audiomassenspeicher oder die Musikdatenbank „Musikbrainz“), oder die Schlagzeilen der Nachrichten (aus DigAS[133]-System) können angezeigt werden. Beispiele sind in Abbildung 8-2 dargestellt.

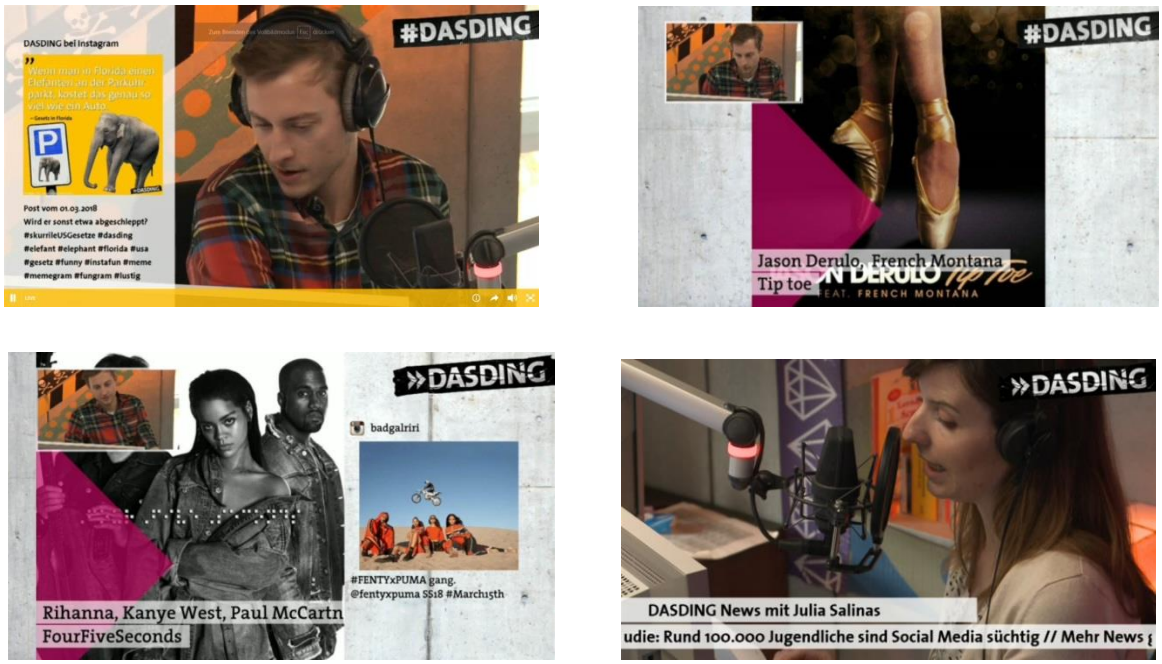


Abbildung 8-2: Kameraeinstellungen des Visual Radio mit Einblendungen

[133] DigAs: Digitaler Aktualitätenspeicher. Leistungsstarker Server für die Speicherung aktueller Programmbeiträge. Ein einfaches Schnittsystem ist ebenfalls integriert.

Die Einblendungen werden mittels XML- oder JSON-Feeds oder Flashs realisiert. Insgesamt sind bis zu 18 Quellen möglich. Für die Einblendungen können im Audio Director Regeln und Bedingungen programmiert werden, um einen reibungslosen Ablauf zu ermöglichen. Danach arbeitet das System automatisch. Radio, Social Media und Video werden so in einer ganz neuen Form miteinander verbunden.

Welcher Moderator oder Gast gerade spricht, erkennt das System anhand des eingestellten Thresholds an jedem Mikrofon. Die Zuordnung welche Kamera auf welchen Sprecher gerichtet wird, muss vorher definiert werden. Danach schneidet die Kamera selbstständig immer auf den Sprecher, der gerade spricht. Sprechen die anwesenden Moderatoren und Gäste gleichzeitig (liegen also beide Signale über dem Threshold) wird durch eingespeicherte Prioritäten geregelt, auf welche Kamera geschnitten wird.

Für ein gutes Ergebnis müssen die Kameras im Vorfeld optimal platziert werden. Des Weiteren gibt es die Möglichkeit die Kameras (wenn sie es zulassen) mittels der Software auch während der Sendung neu auszurichten.

Das Streaming auf die Website verläuft nicht über die Chyronhego-Software. Die Software gibt am Ende lediglich ein Videofile via SDI mit Audio aus. Dieses wird über einen externen Streaming-Server auf die entsprechenden Plattformen weitergeleitet.

DASDING wird aus den beiden Studios DI7 und DI8 im Haus des Hörfunks (HDHF) in Baden-Baden gesendet. In beiden Studios ist das Visual Radio Setting mit drei festinstallierten Kameras und einer mobilen Kamera konfiguriert. Neben dem Live-Betrieb kann das Setup auch zum Produzieren verwendet werden.

Mit dem Visual Radio ergibt sich für die junge Zielgruppe ein völlig neues Radioerlebnis. Dadurch können Gesichtsausdrücke des Moderators, der Studiogäste, Unplugged-Sessions mit bekannten Künstlern oder auch Diskussionsrunden in einer neuen Dimension wahrgenommen werden.

Wenn es nach Ralf Baron (Teamleiter DASDING.technik) geht, ist die Entwicklung des Visual Radio noch lange nicht abgeschlossen. Diese soll sich zukünftig in Richtung eines Players bewegen, der „anfassbar“ ist. Damit sollen individuelle Streams genauso möglich sein, wie beispielsweise die Bereitstellung von weiterführenden Informationen. Damit kann das Visual Radio als Erweiterung

des Radiosignals angesehen werden. Eine Verknüpfung des Radiostreams mit einer Spracherkennung und einer Datenbank wäre dabei der erste Schritt. Abbildung 8-3 veranschaulicht die Idee.

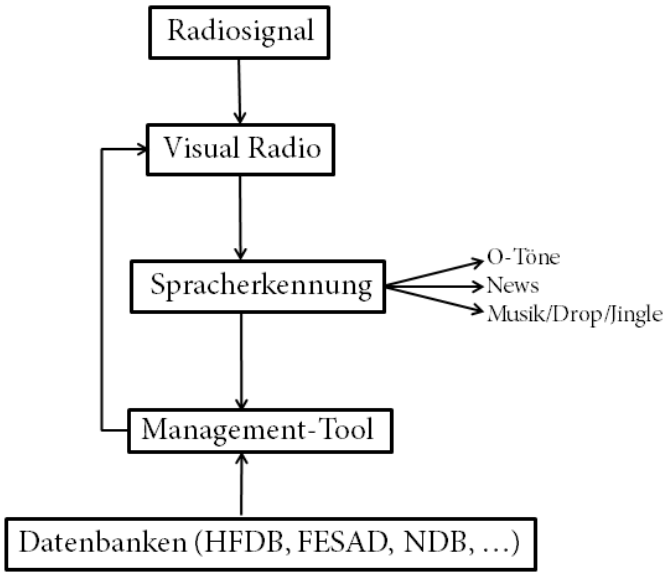


Abbildung 8-3: Schematische Darstellung der Weiterentwicklung des Visual Radio.

Grundlage der Idee ist nach wie vor das Radiosignal. Dieses steuert die Ausgabe des Visual Radio. Allerdings soll das Radiosignal in der Weiterentwicklung mittels einer ARD-weiten Spracherkennung analysiert werden.

Das Radiosignal fungiert als Trigger für die Spracherkennung. Die Spracherkennung muss in der Lage sein zwischen den verschiedenen Elementen (Moderation, Musik, Nachrichten, Jingle, …) zu unterscheiden. Dadurch können sowohl Informationen zu den neuesten Nachrichten oder Moderationen, als auch zum laufenden Titel gesucht werden. Für die Unterscheidung von Musik und Sprache kann auf Technologien des Maschinellen Lernens zurückgegriffen werden. Dafür eignen sich besonders der Entscheidungsbaum (siehe Kapitel 3.1.1), das instanzbasierte Lernen (Kapitel 3.1.2) und das Clustering (Kapitel 3.2.2). Zur Unterscheidung zwischen Sprache und Musik bieten sich als Kriterien die Anzahl der Pausen, der maximale Frequenzbereich, die Amplitude und die Dynamik an. Mithilfe dieser Parameter können Audiodateien hinreichend beschrieben werden.

Das Tool Audio Mining des Fraunhofer – Instituts (siehe Kapitel 6.5) ermöglicht genau das. In Echtzeit kann das Signal in Sprache/Musik/Jingles unterteilt werden. Die Informationen können danach an ein Management-Tool weitergeleitet werden, das die Informationen versteht, aus einer Datenbank filtert und aufbereitet. Das bietet die Weiterentwicklung der NLG-Cloud von AX Semantics. Denn das Unternehmen AX Semantics ist nicht nur im Bereich des Roboterjournalismus

zu finden, sondern seit kurzem auch im Bereich der Chatbots. In diesem Zusammenhang arbeitet AX Semantics mit Alexa- und GoogleHome-Anwendungen. Alexa gibt allerdings nicht die Antwort zurück, sondern wird lediglich zur Spracherkennung genutzt. „Da wo (Alexa) erkannt hat, was du eigentlich von ihr willst, darf sie nichts mehr sagen, sondern gibt ihre Daten über eine API an unsere Plattform weiter. Die darf dann nochmal nachdenken bevor sie antwortet, was Alexa nicht tut."[134] Der AX-Bot muss dabei nicht spezifisch gefragt werden, sondern kann auch eine freier formulierte Frage beantworten.

Für eine Anwendung im Visual Radio muss allerdings keine natürliche Sprache zurückgegeben werden. Schließlich stehen das Radiosignal bzw. thematisch passende Audio- oder Videobeiträge im Mittelpunkt. In diesem Fall würde ein Sprachsignal eine Maschinensprache aufgerufen. „Dann hast du die Situation, dass du mit deiner Stimme die Aktionen eines anderen Systems steuerst."[135] Damit kann beispielsweise die ARD-weite Normdatenbank (NDB) angebunden werden. Der AX-Bot besitzt also alle Informationen, die in dieser Datenbank gespeichert sind und ruft diese bei bestimmten Stichworten auf. Die aufgerufenen Informationen werden aufbereitet und im Visual Radio-Screen eingeblendet.

Somit ist keine komplette Transkription des Audiomaterials nötig. Stattdessen genügt eine Verschlagwortung des Live-Streams. Um die Themenbereiche einzugrenzen, hilft die Stundenuhr. Diese beinhaltet die Grundstruktur einer Sendestunde eines Senders zu einer bestimmten Uhrzeit. Darin wird exakt festgelegt, zu welchen Zeiten Moderationen, Musik und Nachrichten stattzufinden haben. Dadurch wird ein Wiedererkennungsmerkmal für die Konsumenten geschaffen und gleichzeitig die Programmgestaltung vereinfacht. Durch die Vorgaben der Stundenuhr können also Zeitbereiche festgelegt werden, in denen nur bestimmte Informationen bzw. Datenbanken benötigt werden. So kann zum einen Rechenzeit eingespart werden, zum anderen können aber auch die jeweiligen Zeiten akzentuiert bzw. Informationen bei Bedarf auch leichter vorbereitet werden.

Auch eine Google Clips – Anwendung ist in diesem Fall zu überlegen. Die Kamera kann darauf trainiert werden auszulösen, sobald unbekannte Menschen das Studio betreten. Dadurch werden die Studiogäste bzw. Bands direkt beim Eintreffen fotografiert. Dieses Bild kann dann direkt über eine Schnittstelle in die Chyronhego-Software eingespielt werden und auf Facebook/Instagram gepostet werden.

---

[134] Frank Feulner: Interview mit Frank Feulner (AX Semantics). Anhang 3.
[135] ebd.

In welcher Form die gesammelten Informationen zurückgegeben werden, muss wohl überlegt sein. Denn eine Rückgabe in Form eines Textes ist dem Sendeformat nicht in jedem Fall angemessen. Stattdessen ist die Ausspielung in Form eines Videos sehr viel benutzerfreundlicher. Die Plattform Wibbitz (Kapitel 6.6) kann hierbei eine gute Lösung liefern. Bei Wibbitz gibt es die Option, nicht nur Bildmaterial der großen Agenturen sondern auch eigenes Material zu verwenden. Damit kann die Suche in den ARD-Datenbanken via AX Semantics realisiert werden. Durch eine Anbindung an das Portal Wibbitz können daraus die Videos erstellt und direkt eingespielt werden.

Auch IBM Watson Media – Watson Video Enrichment bietet auf dem Gebiet der Videoanalyse spannende Tools an. Mit „Content Search & Discovery" kann das vorhandene Material analysiert und mit Metadaten angereichert werden. Das Material wird besser wiedergefunden und gezielter eingesetzt. Die Videos können auch mit dem Tool „Cognitive Editing" durchsucht werden. Durch die angereicherten Metadaten wird bei der Suche der Daten viel Zeit eingespart.[136]

Auch ein Highlight-Schnitt einzelner Sendungen (beispielsweise der Morningshow) ist mit IBM „Watson – Cognitive Highlight" möglich. Dieser wird automatisch aus dem vorhandenen Kameramaterial geschnitten und nach Sendungsende auf der DASDING-Homepage eingebunden. Bei Gewinnspielen oder besonderen Events können so die Zuschauer auch nachträglich das passende Bildmaterial anschauen und so Teil der Community werden.

Interessant in diesem Zusammenhang ist auch die zukünftige Kommunikationsplattform des Visual Radio. Mit der Möglichkeit des „Amazon Echo Show" (Abbildung 8-4), einer Kombination aus einem Tablet und den Amazon Echo-Varianten, stehen dem Benutzer zum einen die gewohnten Sprachbefehle zur Verfügung. Die Ergebnisse können nun allerdings auch in Form von Bildinhalten angezeigt werden.

Abbildung 8-4: Amazon Echo Show

(Quelle: https://www.amazon.de/Echo-Show-Schwarz-Jetzt-Zeigt-Alexa-Richtig-Viele-Dinge-/dp/B01KGEW44Y)

Ein ähnliches Prinzip bietet sich für die „anfassbare" Variante des Visual Radio an. „Anfassbar" heißt in diesem Zusammenhang eine Bedienung nicht nur via Touchscreen, son-

[136] Vgl. „IBM Watson Media – Watson Video Enrichment", unter: https://www-01.ibm.com/common/ssi/cgi-bin/ssialias?htmlfid=MED12347USEN (Stand: 30.01.2018).

dern auch via Sprache. Kann zusätzlich zum Radiolivestream auch mit der Software gesprochen werden, eröffnet sich eine neue Art der Kommunikation mit der Zielgruppe.

Ändert sich über die Sprachsteuerung auch die Informationsanzeige des Visual Radio, kann sich der Zuhörer eine zusätzliche Informationsquelle erschließen. Abhängig davon, welche Datenbanken angeschlossen sind, können die Themenbereiche breit abgedeckt werden.

Dabei läuft der Radiostream im Hintergrund. Bei der Version über einen Touchscreen können über die Felder mit den Einblendungen (Facebook/Instagram/Nachrichten/Songs) Frames gelegt werden. Diese können bei Interesse von den Nutzern berührt werden. Daraufhin erhalten sie zusätzliche Informationen. Der Umfang der Informationen hängt von der angebundenen Datenbank ab. Die Spracheeingabe des Users ist dabei „höher" priorisiert, als das Radiosignal. Demzufolge muss dieses sowohl akustisch, als auch visuell im Stream in den Hintergrund treten. Stattdessen steht in diesem Moment die Interaktion des Users im Vordergrund.

Die oben beschriebenen Tools und Anwendungen können als Beispiele angesehen werden, die Technologie „Visual Radio" weiterzuentwickeln und dem Hörer so eine wirkliche Ergänzung zu bieten. Die Tools sind allesamt schon veröffentlicht und müssen vor dem Einsatz nur an die jeweiligen Bedingungen und Anforderungen angepasst werden.

Als Ergänzung hierzu bietet sich ein Hackathon an. Dazu werden Softwareentwickler und Programmierer zu einer ein- oder mehrtägigen Veranstaltung eingeladen. Ziel der Veranstaltung ist es, für das Unternehmen Softwareprogramme herzustellen. Dabei steht die Arbeit in Projektteams im Vordergrund. Diese Teams bestehen häufig aus Personen mit unterschiedlichen beruflichen Backgrounds. Mit Hilfe der Hackathons können so Ideen innerhalb kurzer Zeit entwickelt werden, die direkt umgesetzt werden und so als Testobjekte zur Verfügung stehen.

9 Fazit

Künstliche Intelligenz ist eine Thematik, die in den nächsten Jahren in einer Vielzahl von Einsatzgebieten Anwendung finden wird. Durch weiterführende Forschungen mit neuen und mehr Datensätzen werden sich die Systeme immer weiter verbessern und so ein hohes Maß an Sicherheit und Akzeptanz erwerben.

Trotzdem muss jedem Anwender bewusst sein, dass es sich trotz allem nicht um ein System mit menschlichem Verstand handelt. Auch wenn die Leistungsfähigkeit des Menschen in einigen Fällen von der Maschine übertroffen wird, bietet die KI längst nicht in allen Einsatzbereichen Vorteile. Die Grundkompetenz künstlicher Intelligenz liegt in der Geschwindigkeit und der Präzision, mit der die Arbeitsaufträge bearbeitet werden. Allerdings basiert jedes Ergebnis einer KI auf Daten, die vorher von der Software gelernt wurden. Damit sind Eigenschaften wie selbständiges Denken und Handeln, Neues entdecken und Individualität in einer KI nicht vorhanden. Besonders im Bereich der Kreativität ist deshalb ein Einsatz von KI zu überdenken. Auch wenn KIs wie im Fall von Xiaoice[137], in der Lage sind, einen Gedichtband zu schreiben, oder Musik zu komponieren, sind sie dennoch auf menschliche Unterstützung angewiesen. Denn die Stärken der Systeme liegen in den meisten Fällen nur in einem speziellen Bereich. Ein Allrounder, der alle menschlichen Fähigkeiten in derselben Qualität nachahmen kann, konnte bis jetzt noch nicht entwickelt werden.

Vor dem Einsatz einer künstlichen Intelligenz gilt es zudem, genauestens abzuwägen, in welchen Bereichen eine Investition überhaupt sinnvoll ist. Die Stärke und Macht von KI ist zu groß, als dass sie nur um des Einsatzwillens benutzt werden sollte. Der Mehrwert, der sich durch die Anwendung von KI ergibt, stellt sich meist in Form von Zeit- und Kostenersparnissen ein.

Die Entwicklungen im Bereich der KI sind bei Weitem noch nicht abgeschlossen. Weitere Fortschritte werden folgen; Anwendungsgebiete, die bis jetzt noch gar nicht absehbar sind, werden beeinflusst. Im Bereich der Spracherkennung ist die Implementierung der deutschen Sprache nur noch eine Frage der Zeit. Das Deutsche ist aufgrund der komplexen Grammatik durch ständig wechselnde Wortendungen im Bereich der Computerlinguistik schwer in Algorithmen darzustellen. Jedoch wird dieses Problem durch die Verbesserung des maschinellen Lernens und der neuronalen Netze bald gelöst werden.

---

[137] Ein chinesisch sprechender Chatbot (weiblich, 17 Jahre, Kommunikation via Text oder Sprache). Zu Deutsch „kleines Eis“.

Auch der Fehlerquotient wird immer weiter sinken. Dadurch wird die Erkennungssicherheit gesteigert. Handelt es sich um Studioaufnahmen bzw. Sprache ohne Hintergrundgeräusche, erzielt die Spracherkennung bereits jetzt hohe Werte. Geht es allerdings um Audiofiles mit einem niedrigen bis mittleren Grad an Hintergrundgeräuschen bzw. Musik, nimmt die Fehlerrate extrem zu. Deshalb wird sich vor allem in diesem Bereich noch einiges entwickeln. Dann steht auch einer Anwendung im Radioalltag nichts mehr im Wege.

Dass sich die Arbeitswelt dadurch verändern wird, steht außer Frage. Was aber von Wissenschaftlern unterschiedlich vorhergesagt wird, sind die Auswirkungen. Steigt durch den Einsatz von Robotern die Anzahl der Arbeitsplätze an oder nimmt die Summe eher ab? Werden eher Spezialisten gesucht oder Allrounder? In welchen Bereichen ist der Einsatz sinnvoll und welcher Arbeitsplatz kann niemals von einem Roboter übernommen werden? In einer Studie von Carl Benedikt Frey und Michael A. Osborne beschreiben die Autoren die Veränderung der Arbeitswelt aufgrund der Digitalisierung. Grundlage dafür bilden Experteneinschätzungen für 702 Berufe aus den USA. Für diese 702 Berufe liegen Tätigkeitsbeschreibungen des amerikanischen Arbeitsministeriums vor. Ausgehend von diesen Beschreibungen wurde von den Wissenschaftlern nach „Augenmaß“ bestimmt, bis zu welchem Grad die Automatisierung in dem jeweiligen Beruf eingesetzt werden kann. Während sich Physiotherapeuten und Sozialarbeiter keine Sorgen machen müssen, kann im Bereich der Telefonwerbung und der Recherche die Automatisierung ein Großteil der Arbeitsplätze durch Roboter ersetzt werden. Insgesamt könnten laut der Studie in den USA 47% der Arbeitsplätze wegfallen.[138]

Holger Bonin vom Zentrum für europäische Wirtschaftsforschung GmbH in Mannheim hat diese Ergebnisse auf Deutschland übertragen. Für ihn und sein Projektteam ist die Unterscheidung zwischen Beruf und Tätigkeitsfeld entscheidend. Demnach wären nicht ganze Berufe, sondern vielmehr einzelne Tätigkeiten von der Automatisierung betroffen. „Berücksichtigt man dies, so sind in Deutschland nur 12 % der Beschäftigten durch Automatisierung betroffen. Geringqualifizierte und geringverdienende Beschäftigte sind durch die Automatisierung stärker gefährdet.“[139] Es darf allerdings nicht außer Acht gelassen werden, dass die Ergebnisse auf Expertenmeinungen beruhen.

---

[138] Carl Benedikt Frey; Michael A Osborne: The future of employment: How susceptible are jobs to computerization? Oxford 2013, unter: https://www.oxfordmartin.ox.ac.uk/downloads/academic/The_Future_of_Employment.pdf (Stand: 29.01.2018).

[139] Holger Bonin: Übertragung der Studie von Frey/Osborne (2913) auf Deutschland. Mannheim 2015, unter: ftp://ftp.zew.de/pub/zew-docs/gutachten/Kurzexpertise_BMAS_ZEW2015.pdf (Stand: 29.01.2018).

Diese neigen oft dazu, technische Fortschritte zu überschätzen und die Hürden bei der Einführung zu unterschätzen. Hinzu kommt die Tatsache, dass die Veränderung der Arbeitswelt nicht zwangsläufig einen Wegfall der Arbeitsplätze mit sich bringt. Stattdessen handelt es sich vielmehr um eine Veränderung. Denn zum einen können nicht alle Arbeitsschritte eines Berufs automatisiert werden. Zum anderen bringt die Automatisierung eine erhöhte Störanfälligkeit mit sich. Das bedeutet, dass durch den Fortschritt nicht zwangsläufig weniger Arbeitskräfte benötigt werden, sondern vielmehr besser ausgebildetes Personal gebraucht wird. Dieses Personal muss sich trotz der Automatisierung mit den technischen Arbeitsschritten auskennen, um möglicherweise auftretende Probleme zu beheben. Dazu ist ein hohes Maß an Erfahrung notwendig.[140] „Facharbeiter haben sie, Roboter haben sie nicht."[141]

Nichtsdestotrotz ist ein Einsatz von KI ein Weg in die Zukunft. Auch wenn die Umstellung bzw. das Eingestehen dieser Möglichkeiten nicht immer leicht ist, überwiegen die Vorteile. Diese wahrzunehmen und umzusetzen ist dabei die Hauptaufgabe. Schließlich ist die Sprache das Übertragungsmedium eines Rundfunksenders. Die Weitergabe von Informationen via natürliche Sprache ist die schnellste Form der Übermittlung an die Zuhörerschaft. Warum dann nicht auch die Datenbanken dahinter mit einer Sprachsuche ausstatten, statt weiter auf die reine Wortsuche in Texten zu bauen? Denn passendes Audiomaterial kann nur mit einer Suche innerhalb des Audiomaterials schnell bewerkstelligt werden bzw. mit einer Transkription des Audiomaterials in Echtzeit, um anhand der Metadaten suchen zu können. Damit gibt es keine Barriere zwischen der natürlichen Sprache und der Arbeit mit Texten. Genau in dieser Art des Arbeitens liegt die Zukunft.

---

[140] Vgl. Holger Bonin: Übertragung der Studie von Frey/Osborne (2913) auf Deutschland. Mannheim 2015, unter: ftp://ftp.zew.de/pub/zew-docs/gutachten/Kurzexpertise_BMAS_ZEW2015.pdf (Stand: 29.01.2018).

[141] Max Rauner: Die Pi-mal-Daumen-Studie. 2017, unter: http://www.zeit.de/2017/11/kuenstliche-intelligenz-arbeitsmarkt-jobs-roboter-arbeitsplaetze/komplettansicht (Stand: 29.01.2018).

Literaturverzeichnis

Agrawal, R.; Imielinski, T.; Swami, A. (1993): Database mining. A performance perspective. In: *IEEE Trans. Knowl. Data Eng.* 5 (6), S. 914–925. DOI: 10.1109/69.250074.

Ajiboye, A. Bolu; Willett, Francis R.; Young, Daniel R.; Memberg, William D.; Murphy, Brian A.; Miller, Jonathan P. et al. (2017): Restoration of reaching and grasping movements through brain-controlled muscle stimulation in a person with tetraplegia. A proof-of-concept demonstration. In: *The Lancet* 389 (10081), S. 1821–1830.

Banko, Michele; Brill, Eric: Scaling to Very Very Large Corpora for Natural Language Disambiguation. Online verfügbar unter http://www.aclweb.org/anthology/P01-1005.

Bonin, Holger; Terry, Gregory; Zierahn, Ulrich (2015): Übertragung der Studie von Frey/Osborne (2913) auf Deutschland. Zentrum für Europäische Wirtschaftsforschung GmbH. Mannheim (Kurzexperise Nr. 57), zuletzt geprüft am 29.01.2018.

Carstensen, Kai-Uwe; Ebert, Christian; Ebert, Cornelia; Jekat, Susanne J.; Klabunde, Ralf; Langer, Hagen (Hg.) (2010): Computerlinguistik und Sprachtechnologie. Eine Einführung. 3., überarb. und erw. Aufl. Heidelberg: Spektrum Akad. Verl. (Spektrum Lehrbuch).

Chow, Yen-Lu; Schwartz, Richard (1990): The N-Best Algorithm: Efficient Procedure for Finding Top N Sentence Hypotheses. In: *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP-90)*, S. 81–84.

Eberl, Ulrich (2011): Zukunft 2050. Wie wir schon heute die Zukunft erfinden. Weinheim: Beltz & Gelberg.

Englert, Fred; Euler, Stephan; Wolf, Dietrich (1992): Zur Variabilität sprachlicher Äußerungen in der sprecherunabhängigen Einzelworterkennung. In: Helmut Mangold (Hg.): Sprachliche Mensch-Maschine-Kommunikation, Bd. 31. München: Oldenbourg.

Ertel, Wolfgang (2016): Grundkurs Künstliche Intelligenz. Eine praxisorientierte Einführung. 4., überarbeitete Auflage. Wiesbaden: Springer Vieweg (Computational Intelligence).

Euler, Stephen (2006): Grundkurs Spracherkennung. Vom Sprachsignal zum Dialog - Grundlagen und Anwendung verstehen - Mit praktischen Übungen. 1. Aufl. Wiesbaden: Friedr. Vieweg & Sohn Verlag/GWV Fachverlage GmbH Wiesbaden (Computational Intelligence).

Fant, Gunnar (1960): Acoustic Theory of Speech Production. Den Haag: mouton.

Ferrucci, David; Brown, Eric; Chu-Carroll, Jennifer; Fan, James; Gondek, David; Kalyanpur, Aditya A. et al. (2010): Building Watson. An Overview of the DeepQA Project. In: *AIMag* 31 (3), S. 59–79. DOI: 10.1609/aimag.v31i3.2303.

Ford, Martin (2016): Aufstieg der Roboter. Wie unsere Arbeitswelt gerade auf den Kopf gestellt wird - und wie wir darauf reagieren müssen. Kulmbach: Plassen Verlag.

Frey, Carl Benedikt; Osborne, Michael A. (2017): The future of employment. How susceptible are jobs to computerisation? In: *Technological Forecasting and Social Change* 114, S. 254–280. DOI: 10.1016/j.techfore.2016.08.019.

Görz, Günther; Schneeberger, Josef; Schmid, Ute (Hg.) (2014): Handbuch der künstlichen Intelligenz. 5., überarb. und aktualisierte Aufl. München: Oldenburg.

Graefe, Andreas (2016): Guide to Automated Journalism. Tow Center for Digital Journalism, Columbia Journalism School. New York City.

Haim, Mario; Graefe, Andreas (2017): Automated News: Better than expected? In: *Digital Journalism* 5 (8), S. 1044–1059. DOI: 10.1080/21670811.2017.1345643.

Hays, James; Efros, Alexei A. (2007): Scene completion using millions of photographs. In: *Commun. ACM* (26), S. 87–94. DOI: 10.1145/1400181.1400202.

Hoffmann, Norbert (1993): Kleines Handbuch Neuronale Netze. Anwendungsorientiertes Wissen zum Lernen und Nachschlagen. Wiesbaden: Vieweg+Teubner Verlag. Online verfügbar unter http://dx.doi.org/10.1007/978-3-322-91565-8.

Jones, Nicola (2014): Computer science. The learning machines. In: *Nature* 505 (7482), S. 146–148. DOI: 10.1038/505146a.

Karger, Reinhard (2017): Künstliche Intelligenz: Keine Angst vor Roboter-Reportern, 20.02.2017. Online verfügbar unter http://www.zeit.de/karriere/2017-02/kuenstliche-intelligenz-journalismus-roboter-digitalisierung, zuletzt geprüft am 21.08.2017.

Kuhlen, Rainer; Semar, Wolfgang; Strauch, Dietmar (Hg.) (2013): Grundlagen der praktischen Information und Dokumentation. Berlin: de Gruyter (Handbuch zur Einführung in die Informationswissenschaft und -praxis).

Lang, Andreas; Holley, Marcus; Dittmann, Jana (2004): StirMark for Audio: Unterschiede zwischen Musik und Sprache. In: Klaus Peter Fähnrich (Hg.): Von e-Learning bis e-Payment 2004. Das Internet als sicherer Marktplatz ; Tagungsband LIT '04, 29. September - 1. Oktober 2004, Leipzig. Berlin: Akad. Verl.-Ges. Aka. Online verfügbar unter http://www.fitl.htwk-leipzig.de/T/A/a/13_LIT04_Lang.pdf, zuletzt geprüft am 23.01.2018.

Levenshtein, Vladimir (1966): Binary codes capable of correcting deletions, insertions, and reversals. In: *Soviet Physics Doklady* 10 (8), S. 707–710. Online verfügbar unter https://nymity.ch/sybilhunting/pdf/Levenshtein1966a.pdf.

Lighthill, James (1973): Artificial Intelligence: A General Survey. In: Science Research Council (Hg.): Artificial Intelligence. A Paper Symposium: Science Research Council, S. 3–18. Online verfügbar unter http://www.chilton-computing.org.uk/inf/literature/reports/lighthill_report/p001.htm, zuletzt geprüft am 01.02.2018.

Loughran, Michael; Zimmerman, Phil: About IBM Watson – Factsheet. Hg. v. IBM. Online verfügbar unter http://www-07.ibm.com/systems/hk/power/news/pdf/IBM_Watson_Fact_Sheet.pdf.

Ludwig, Bernd (2013): Spracherkennung. In: Rainer Kuhlen, Wolfgang Semar und Dietmar Strauch (Hg.): Grundlagen der praktischen Information und Dokumentation. Berlin: de Gruyter (Handbuch zur Einführung in die Informationswissenschaft und -praxis), S. 313–321.

Lundberg, George A. (1938): Primary Mental Abilities. L. L. Thurstone. In: *American Journal of Sociology* 44 (2), S. 310–311. DOI: 10.1086/217986.

Mainzer, Klaus (2016): Künstliche Intelligenz - Wann übernehmen die Maschienen? // Künstliche Intelligenz - Wann übernehmen die Maschinen? Berlin, Heidelberg: Springer Berlin Heidelberg; Springer (Technik im Fokus).

Mangold, Helmut (Hg.) (1992): Sprachliche Mensch-Maschine-Kommunikation. München: Oldenbourg.

McCulloch, Warren S.; Pitts, Walter (1943): A logical calculus of the ideas immanent in nervous activity. In: *Bulletin of Mathematical Biophysics* 5 (4), S. 115–133. DOI: 10.1007/BF02478259.

Menzel, Wolfgang (2014): Sprachverarbeitung. In: Günther Görz, Josef Schneeberger und Ute Schmid (Hg.): Handbuch der künstlichen Intelligenz. 5., überarb. und aktualisierte Aufl. München: Oldenburg, S. 473–526.

Norvig, Peter; Russell, Stuart (2012): Künstliche Intelligenz, 3.Auflage. München, [s.l.]: Pearson Deutschland; Pearson Studium (Pearson Studium - IT).

Patterson, Dan W. (1997): Künstliche neuronale Netze. Das Lehrbuch. 2. Aufl. Haar bei München: Prentice Hall.

Pfister, Beat; Kaufmann, Tobias (2017): Sprachverarbeitung - Grundlagen und Methoden der Sprachsynthese und Spracherkennung // Sprachverarbeitung. Grundlagen und Methoden der Sprachsynthese und Spracherkennung. Berlin, Heidelberg, s.l.: Springer Berlin Heidelberg.

Reetz, Henning (2003): Artikulatorische und akustische Phonetik. 2., verb. Aufl. Trier: WVT Wiss. Verl. Trier.

Rich, Elaine; Knight, Kevin; Nair, Shivashankar B. (2009): Artificial intelligence. 3rd ed. Neh Delhi: Tata McGraw-Hill.

Russell, Stuart; Norvig, Peter; Kirchner, Frank (2012): Künstliche Intelligenz. Ein moderner Ansatz. 3., aktualisierte Aufl. München: Pearson Higher Education (Always learning).

Slotta, Jörn (2012): Vergleich von Algorithmen zur Assoziationsanalyse basierend auf Webserver LogVles. Bachelor-thesis. HAW Hamburg, Hamburg. Department Informatik. Online verfügbar unter http://edoc.sub.uni-hamburg.de/haw/volltexte/2012/1747/pdf/BA_Slotta.pdf.

Spearman, Charles (1904): "General Intelligence," Objectively Determined and Measured. In: *The American Journal of Psychology* 15 (2), S. 201. DOI: 10.2307/1412107.

Theuerkauf, Judith (2012): Schreiben im Ingenieurstudium. Paderborn: Schöningh (effektiv und effizient zur Bachelor-, Master- und Doktorarbeit).

Utz, Wolfgang (2017): SWR - Zahlen Daten Fakten. Südwestrundfunk. Stuttgart. Online verfügbar unter https://www.swr.de/-/id=9391984/property=download/nid=7687068/9bwsth/index.pdf.

van den BERG, J. (1956): Direct and indirect determination of the mean subglottic pressure; sound level, mean subglottic pressure, mean air flow, subglottic power and efficiency of a male voice for the vowel (a). In: *Folia phoniatrica* 8 (1), S. 1–24.

Wahlster, Wolfgang (Hg.) (2000): Verbmobil. Foundations of speech-to-speech translation ; with 88 tables. Berlin: Springer (Artificial intelligence).

Wrobel; Stefan, Joachims; Thorsten, Morik; Katharina (2014): Maschinelles Lernen und Data Mining. In: Günther Görz, Josef Schneeberger und Ute Schmid (Hg.): Handbuch der künstlichen Intelligenz. 5., überarb. und aktualisierte Aufl. München: Oldenburg, S. 405–471.

Zaun, Detlef Peter (2016): Künstliche neuronale Netze und Computerlinguistik. Reprint 2016. Tübingen: Max Niemeyer Verlag (Linguistische Arbeiten, 406).

Zemlin, Willard R. (op. 1998): Speech and hearing science. Anatomy and physioloy. 4.ed. Boston: Allyn and Bacon.

Zöller-Greer, Peter (2010): Künstliche Intelligenz. Grundlagen und Anwendungen ; mit einer Einführung in die Logik [Aussagen- und Prädikatenlogik], PROLOG, Expertensysteme, Fuzzy-Systeme, neuronale Netze und genetische Algorithmen. 2. Aufl. Wächtersbach: Composia-Verl. (Reihe Wissen & Praxis kompakt).

Quellenverzeichnis

Bilder für Bilderkennungstest (Visual Recognition / Maschinelles Sehen):

Brad Pitt:

http://www.observerbd.com/2017/11/27/1511791440.jpg

https://i2-prod.mirror.co.uk/incoming/article189804.ece/ALTERNATES/s298/Brad-Pitt.gif

https://upload.wikimedia.org/wikipedia/commons/5/51/Brad_Pitt_Fury_2014.jpg

Leonardo DiCaprio:

https://upload.wikimedia.org/wikipedia/commons/thumb/9/9a/Leonardo_DiCaprio_2010.jpg/220px-Leonardo_DiCaprio_2010.jpg

https://cdn1.thr.com/sites/default/files/imagecache/landscape_928x523/2014/11/leonardo_dicaprio.jpg

http://d17zbv0kd7tyek.cloudfront.net/wp-content/uploads/2015/06/leonardo-dicaprio-fb.jpg

Angelina Jolie:

https://peopledotcom.files.wordpress.com/2017/06/angelina-jolie-7.jpg

https://image.gala.de/20509790/large1x1-460-460/3a1bb5e3593f5cad714d2061dc8e05f3/nK/angelina-jolie-starportraet--5991905-.jpg

https://upload.wikimedia.org/wikipedia/commons/thumb/8/83/Angelina_Jolie_at_the_launch_of_the_UK_initiative_on_preventing_sexual_violence_in_conflict%2C_29_May_2012_%28cropped%29.jpg/220px-Angelina_Jolie_at_the_launch_of_the_UK_initiative_on_preventing_sexual_violence_in_conflict%2C_29_May_2012_%28cropped%29.jpg

Tom Cruise:

https://pixel.nymag.com/imgs/daily/vulture/2017/06/14/14-tom-cruise.w710.h473.2x.jpg

https://www.biography.com/.image/t_share/MTE5NDg0MDU0OTM2NTg1NzQz/tom-cruise-9262645-1-402.jpg

https://image.gala.de/20511760/large1x1-460-460/ba22ba50543964a177f48d7861d55049/Cf/tom-cruise-ge--6411731-.jpg

Heidi Klum:

https://image.gala.de/20514732/large1x1-460-460/9ab73961ac248825d910b025723f227f/GS/heidi-klum--6271749-.jpg

https://media.glamour.com/photos/5696ce4401ed531c6f006feb/master/pass/beauty-2015-08-heidi-klum-trump-response-main.jpg

https://upload.wikimedia.org/wikipedia/commons/thumb/6/60/Heidi_Klum_Judges_Red_Carpet_event_April_2014_%28cropped%29.jpg/1200px-Heidi_Klum_Judges_Red_Carpet_event_April_2014_%28cropped%29.jpg

Matthias Schweighöfer:

https://image.gala.de/20481828/large1x1-460-460/b0a602eceaed10a9019bf52bcaf4b94f/pr/matthias-schweighoefer--6403205-.jpg

https://assets.cdn.moviepilot.de/files/c6cf07855301f44a0fc2f7921891edaeb119b6ef50b2489b7b504379e137/matthias-schweighoefer_.jpg

J

https://image.stern.de/7207506/16x9-940-529/98746c219df92dd09baa2deca04089e8/wv/matthias-schweighoefer.jpg

Til Schweiger:

https://www.welt.de/img/vermischtes/mobile152547941/6622508407-ci102l-w1024/Til-Schweiger.jpg

https://i0.web.de/image/490/31755490,pd=2/til-schweiger.jpg

http://barefootfilms.de/assets/Team/_resampled/image-Tilret.jpg

Heike Makatsch:

http://www.tagesspiegel.de/images/heike-makatsch/13370692/1-format43.jpg

https://assets.cdn.moviepilot.de/files/7971f297b4be9b589a5a88b76a41235091c912a0c30ade94140d8ca32928/heike_makatsch_cut.jpg

https://image.gala.de/20503694/large1x1-460-460/7037941bbb8d10ac8f3085b772c6637/FU/heike-makatsch--6283141-.jpg

Veronica Ferres:

https://image.stern.de/6391156/uncropped-0-0/be1bf879054bfec4c171bb506c59ff5/Jm/veronica-ferres.jpg

https://www.welt.de/img/vermischtes/mobile138224624/6862500767-ci102l-w1024/ONLINE-2.jpg

https://image.gala.de/20328984/uncropped-0-0/1585158158aab32b16083c4d2fe17a66/bF/veronica-ferres--10891001-.jpg

Nora Tschirner:

https://www.welt.de/img/vermischtes/mobile154606293/2781353387-ci16x9-w1200/Nora-Tschirner-spricht-Lara-Croft.jpg

https://programm.ard.de/files/265/2007/2698/3019/9254/00_Tschirner_Tatort_Tschirner_Ulmen_Weimar_1408.jpg

http://www.schauspielervideos.de/FM/fotos/21/21545/nora-tschirner-kf1113914_gr_1024.jpg

Alle Bilderquellen für diese Testreihe wurden am 21.12.2018 abgerufen.

Anhang

1. IPA-Zeichen[142]

Vokale:

ʔi:	[ʔi:gəl]	Igel	ɔ	[fɔlk]	Volk
ʔI	[ʔInzəl]	Insel	ɔ	[bɔk]	Bock
i:	[ʃmi:d]	Schmied	u:	[blu:mə]	Blume
i	[politIk]	Politik	u:	[mu:sə]	Muse
e:	[ze:]	See	u	[musik]	Musik
e	[metal]	Metall	U	[bUs]	Bus
e:	[be:t]	Beet	Y	[mYl]	Müll
ɛ	[bɛt]	Bett	y:	[ky:l]	kühl
ɛ:	[kɛ:zə]	Käse	y:	[hy:tə]	Hüte
a	[dax]	Dach	Y	[hYtə]	Hütte
a	[kan]	kann	y	[kymogram]	Kymogramm
a:	[ka:n]	Kahn			
a:	[ta:k]	Tag	ə	[ʃu:lə]	Schule
ø:	[ʔø:l]	Öl	ə	[lesən]	lesen
ø:	[hø:lə]	Höhle			
œ	[hœlə]	Hölle	ɐ	[lesɐ]	Leser
œ	[kœln]	Köln	ɐ	[mUtɐ]	Mutter
ø	[ʔøkono:m]	Ökonom			
o:	[ko:l]	Kohl	aI	[fraI]	frei
o:	[bo:g]	bog	aU	[fRaU]	Frau
o	[politIk]	Politik	ɔI	[nɔI]	neu

Konsonanten:

p	[plan]	Plan	ʒ	[gaRa:ʒə]	Garage
b	[blu:mə]	Blume	ç	[ʔIç]	ich
t	[ta:k]	Tag	x	[ʔaxt]	acht
d	[dax]	Dach	ʁ	[ʁa:m]	Rahm *(frikatives* [R]*)*
k	[ko:l]	Kohl	h	[haUs]	Haus
g	[gla:s]	Glas	m	[maUs]	Maus
ʔ	[ʔana]	Anna	n	[naxt]	Nacht
f	[fo:gəl]	Vogel	ŋ	[vaŋə]	Wange
v	[vɛlə]	Welle	l	[lIçt]	Licht
s	[bus]	Bus	r	[ra:m]	Rahm *(Zungen-*R*)*
z	[ze:]	See	R	[Ra:m]	Rahm *(Zäpfchen-*R*)*
ʃ	[ʃu:le]	Schule	j	[ja:kt]	Jagd

[142] Deutsche Phonetik, unter: http://www.obib.de/Schriften/Phonetik/deutsch/deutsch.php (Stand: 31.10.2017).

2. Technische Daten Smartphones

a. Huawei Mate 10 Pro (Auszug)[143]

Größe: Breite: 74,5 mm Höhe: 154,2 mm
Tiefe: 7,9 mm Gewicht: ca. 178 g

Display:

- 6.0" HUAWEI FullView Display
- 18:9 OLED Display
- 2160 x 1080 Pixel, 402 ppi
- 16,7 Mio. Farben, Farbsättigung (NTSC): 112 %
- Hoher Kontrast: 70.000:1 (Typischer Wert)

Chipsatz

- HUAWEI Kirin 970
- Achtkern-CPU (4*Cortex A73 2,36 GHz + 4*Cortex A53 1,8 GHz) + i7 Co-Prozessor,
- Mali-G72 MP12 GPU
- NPU: Neural Network Processing Unit

Betriebssystem: Android™ 8.0

Speicher: 6 GB RAM +128 GB ROM

Dual-SIM-Unterstützung: BLA-L09: Single-SIM BLA-L29: Dual-SIM

Konnektivität:

- WLAN 2.4G/5G, 802.11 a/b/g/n/ac mit Unterstützung für Wi-Fi Direct
- BT 4.2, BLE
- aptX/aptX HD und LDAC HD Audio
- USB Type C
- DisplayPort 1.2

Sensoren: Fingerabdruck, Umgebungslicht, Näherung, Beschleunigung, Kompass, Gyroskop, Barometer, Hall, IrDA

Kamera:

- Hauptkamera: 20 MP Monochrom + 12 MP RGB Leica Dual Kamera
  Summilux-H 1:1.6/27 ASPH, OIS (Optische Bildstabilisierung), BSI CMOS
  PDAF + CAF + Laser + Tiefen-Autofokus
  2x Hybridzoom, 4K Videoaufnahme
- Frontkamera: 8 MP FF, f/2.0

Audio: Audio-Dateiformate: MP3, MP4, 3GP, OGG, AMR, AAC, FLAC, WAV, MIDI

Video: Videodateiformate: 3GP, MP4

EMUI: EMUI 8.0

Akku: 4000 mAh (Typischer Wert)

---

[143] Vgl. http://consumer.huawei.com/de/phones/mate10-pro/specs/ (Stand: 27.11.2017).

b. iPhone X (Auszug)[144]

Größe: Höhe: 143,6 mm Breite: 70,9 mm
Tiefe: 7,7 mm Gewicht: 174 g

Display:

- 5,8" All-Screen OLED Multi-Touch Display
- HDR Display
- Typisches Kontrastverhältnis: 1.000.000:1
- Display mit großem Farbraum (P3)
- Maximale typische Helligkeit: 625 cd/m²
- Super Retina HD Display
- 2436 x 1125 Pixel bei 458 ppi
- True Tone Display
- 3D Touch
- Fettabweisende Beschichtung

Chip:

- A11 Bionic Chip mit 64 Bit-Architektur
- Integrierter M11 Motion Coprozessor
- Neurale Architektur

Kamera:

- 12 Megapixel Kamera mit Weitwinkel- und Teleobjektiv
- Weitwinkel: f/1.8 Blende
- Optischer Zoom, bis zu 10x digitaler Zoom
- Optimiertes lokales Tone Mapping
- Aufgenommene Bildformate: HEIF und JPEG
- Geotagging für Fotos
- Teleobjektiv: f/2.4 Blende
- Rauschunterdrückung
- Körper- und Gesichtserkennung
- Belichtungsregler

Videoaufnahme

- 4K Videoaufnahme mit 24 fps, 30 fps oder 60 fps
- 1080p HD Videoaufnahme mit 30 fps oder 60 fps / 720p HD Videoaufnahme mit 30 fps
- Unterstützung für Zeitlupenvideo in 1080p mit 120 fps oder 240 fps
- Zeitraffervideo mit Bildstabilisierung, Cinematic Videostabilisierung (1080p und 720p)
- Kontinuierlicher Autofokus
- Körper- und Gesichtserkennung, Rauschunterdrückung
- Aufgenommene Videoformate: HEVC und H.264

TrueDepth Kamera

- 7 Megapixel Kamera
- Porträtlicht (Beta)
- 1080p HD Videoaufnahme
- f/2.2 Blende
- Rückwärtig belichteter Sensor
- Automatische Bildstabilisierung
- Belichtungsregler
- Porträtmodus
- Animoji
- Retina Blitz
- Automatisches HDR
- Körper- und Gesichtserkennung
- Serienbildmodus
- Selbstauslöser

Face ID: Ermöglicht durch eine TrueDepth Kamera für Gesichtserkennung

[144] Vgl. https://www.apple.com/de/iphone-x/specs/ (Stand: 21.12.2018)

3. Interview Frank Feulner (AX Semantics)

Das Interview mit Frank Feuler fand am 21.12.2017 zwischen 10.30 Uhr und 11.30 Uhr bei der Firma AX Semantics in Stuttgart in der Marienstraße 23 statt. Frank Feulner ist Mitbegründer von AX Semantics und als Chief Visionary Officer Mitglied der Geschäftsleitung.

Wie kam es zu der Software?

Als wir mit der Entwicklung anfingen, hieß die ganze Firma noch aexea. Wir waren bekannt als Redaktions- und Kommunikationsberatung. Wir haben mehr oder weniger für die DAX-Top 30 Kommunikationsstrategien entwickelt und im Rahmen dieser Kommunikationsstrategien auch Redaktionsdienstleistungen verkauft. Das können Sie sich wirklich so vorstellen: Ein großer deutscher Hersteller für Optiken möchte Verkaufstexte für seine Elektronenmikroskope schreiben. Und aexea schickt mich los. Ich werde zwei Wochen lang bei diesem Hersteller ausgebildet, weiß anschließend alles über Mikrotomie und Mikroskopie und schreibe dann Texte, die diesen Hersteller 300 € pro Seite kosten. Das war der Rahmen, in dem wir uns damals bewegt haben. Das hat bis ungefähr 2008 auch super geklappt. Bis dann sogenannte Crowdsourcing-Anbieter auf den Markt kamen. Firmen wie Textbroker, Textprovider und GreatContent. Die haben gesagt, dass man einen Text nicht mehr mit einem gut ausgebildeten Journalisten schreiben muss, sondern auch mit 30 schlecht ausgebildeten schreiben kann und sich dann die besten Stücke zusammenbrokert. Die haben dann den Claim auf den Markt gedrückt, dass man einen Text auch für 3€ pro Seite kriegt. Das hat die Zahlungsbereitschaft unserer Kunden ausgetrocknet.

Wir haben im Rahmen unserer Beratungstätigkeit auch schon analytische Software entwickelt, beispielsweise um unsere Studien auszuwerten, um Textqualität oder Textmerkmale zu analysieren. Deshalb hatten wir da schon Berührungspunkte. Da haben wir gesagt: Gut, günstig und skalierbar muss es jetzt sein. Das gewährleisten wir durch eine Software, die Texte günstig und skalierbar erzeugt. Den Prototypen haben wir dann in Excelformeln geschrieben.

In welchen Bereichen wird die Software hauptsächlich angewendet?

Als Kernmärkte haben sich bis jetzt drei herausgebildet. Das ist einmal ganz stark der eCommerce. Überall dort, wo Sie einen großen Onlineshop haben, der auch über ein gewisses Maß an Produktdaten verfügt, können Sie Texte, die das Produkt mit dem Kunden diskutieren, also die ihm sagen: Das ist Gore-Tex und ja, das ist wasserdicht und mit dieser Jacke kannst du dich in den Regen stellen. Die diesen Transfer in seine Lebensrealität machen, diese Texte können Sie jetzt auch auf Basis von Daten schreiben. Die Maschine wird trainiert zu erkennen, welches Feature welche Argumentation auslösen sollte. Anschließend erzählt sie dann für jeden einzelnen Artikel auf Basis der Daten eine Geschichte, die dem Kunden hilft, sich für oder gegen das Produkt zu entscheiden. Das macht etwa 80 % unseres Umsatzes aus.

Für die Größenordnung: Wir haben dieses Jahr pro Monat etwa 30 Millionen Stories geschrieben, ca. 80% davon kommen aus dem eCommerce. Dann gibt es noch die anderen 20 %, die zum größten Teil aus den journalistischen Arbeiten kommen. Das sind dann Berichtsformen, die auch da-

tengetrieben sind und vor allem wiederkehrend sind. Also, die sich dadurch auszeichnen, dass sie mehr oder weniger jeden Tag denselben Bericht schreiben nur mit anderen Inhalten. Das sind dann Themen wie Wetter, das ist viel lokaler Sport. Weshalb lokal? Weil man dadurch dass die Texte so unglaublich günstig sind, jetzt plötzlich auch eine tiefere Berichterstattung in die Breite treiben kann. Dass man plötzlich einen Bericht schreiben kann, der jede einzelne Spielminute potenziell dokumentiert und das für eine Begegnung wie den SV Waltershofen gegen den TSC Kißlegg – haben Sie noch nie von gehört. Die Audience ist etwa 1000 Leute groß. Aber Sie würden trotzdem zögern, dort einen Reporter hinzuschicken und genauso zu allen anderen Begegnungen dieser Stufe. Weil das Verhältnis zwischen der Reichweite, die Sie dadurch bekommen und dem Aufwand, den Sie machen müssten, einfach nicht gegeben ist. Mit uns jetzt schon! Und deshalb wird in der Journalistik mit uns jetzt vor allem hyperlokalisierter Content produziert: das Wetter heruntergebrochen auf Postleitzahlen oder Straßenzugebenen. Individualisierter Content ist gerade ein großer Trend. Beispielsweise sagt man: Ich bekomme meine Wetterdaten. Aber die Plattform, die mir das Wetter berichtet, weiß, dass ich Allergiker bin und mit dem Fahrrad zur Arbeit fahre.

Man hat eigentlich immer datengetriebene Formate. Dadurch versucht man, ein bisschen Zeit in den Redaktionen frei zu machen. Wir sehen bei unseren Kunden, dass diese Zeit dann meistens ins Brand-building fließt. Also, dass man gar nicht, wie anfangs befürchtet, seine Journalisten freisetzt, sondern dass man sagt: OK, ihr müsst euch jetzt nicht mehr mit dem Veranstaltungskalender herumprügeln, der wird jetzt automatisiert genauso wie das Wetter und das Horoskop. Was ihr stattdessen für uns tun könnt, ist informativere, neuere Darstellungsformate zu finden, in einer größeren Breite auch Recherche zu betreiben, mehr vor Ort zu gehen und so weiter. Und das ist auch tatsächlich das, was bei so Schlüsselkunden, wie der Nordwest Zeitung oder der Augsburger Allgemeinen, auch gemacht wird.

Dann hat man eigentlich eine Situation, wo sich drei große Trends abzeichnen. Das ist einmal, dass man in Formaten denken kann, die sich in Echtzeit abspielen. Einfach, weil die Maschine unglaublich schnell arbeitet. Die bewegt sich aber auch in Richtung der Individualisierung. Dass man also das Publikum in kleineren Clustern denken kann, weil es auch einfach nicht so viel mehr kostet den Artikel nochmal in einer anderen Färbung zu schreiben.

Und das dritte ist eben tatsächlich, dass man den Content auch intrinsisch mehrsprachig denken kann. Die Maschine spricht mittlerweile 28 oder 29 Sprachen. Die Zahl wird von Interview zu Interview größer. Wir brauchen im Schnitt zwei bis zehn Wochen, um eine Sprache zu entwickeln. Jetzt ist gerade Ungarisch hinzugekommen.

Das bedeutet aber auch, dass Sie jetzt, wenn Sie Teile in der Audience haben, die eine andere Muttersprache haben, die sie jetzt auch in ihrer Muttersprache abholen können. Wir sprechen jetzt zum Beispiel auch Türkisch und Arabisch. Selbst so Klopper wie Finnisch und Ungarisch sind wir jetzt eben dabei. Aus linguistischer Sicht ist das echt eine große Sache.

Das heißt eben auch, dass Sie die Audience nicht mehr in einer konsensuellen Sprache, sondern vielsprachig denken können.

Heißt das ich tippe einen deutschen Text ein und es wird automatisch übersetzt? Oder muss ich die Übersetzung auch mitanlegen?

Sie bringen der Maschine ja immer bei, wie Sie einen Text determinieren. Die Maschine lernt zwei Sachen von Ihnen. Das ist zum Einen, wie man Daten liest und wie man sich in einem bestimmten Thema über diese Daten ausdrückt. Dadurch bauen Sie einen Entscheidungsbaum auf, der immer von der Analyse der Daten herkommt. Dieser Entscheidungsbaum endet irgendwo in echten sprachlichen Strukturen, also da, wo die Maschine Wörter und ihr bekannte Phrasen auswählt, die sie dann zu einem Text zusammenfügt. An dieser Stelle können Sie nochmal verästeln und dann können Sie nochmal Varianten für diese 27 Sprachen für verschiedene Lokalen trainieren. Das ist auch gut so, dass Sie das selber machen müssen, weil Transduktion, also das Übersetzen von einer Ausgangsoberfläche zu einer anderen, leider den Nachteil hat, dass man alle Sprachbilder und alle Gedankenkonzepte mitübernimmt. Das kennen Sie selber: Ministerpräsidenten in Englisch, „We are all sitting in the same boat", … Das sind alles Ergebnisse, die Sie von einer KI eigentlich nicht erwarten. Ein ganz praktisches Beispiel: Fernsehtexte wurden von uns für Spanien in Spanisch aufgesetzt. Hat auch super funktioniert, wurde dann nach Chile lokalisiert und kam mit vernichtendem Feedback zurück: Die spinnen ja die Roboter - der Fernseher wird in das Wohnzimmer gestellt? Was sind denn das für Wilde. Der heilige Ort der Konversation wird plötzlich durch einen Fernseher verseucht. Der gehört doch ins Schlafzimmer. Und spätestens dann müssen Sie auch einwirken, weil Sie das durch logische Transduktion nicht hinkriegen.

Das dritte Anwendungsfeld kristallisiert sich jetzt gerade so heraus und da ist auch mein Business-Development jetzt gerade ganz stark dran, in diese Märkte reinzudrücken. Das ist sozusagen der Funktionaltext. Dort wo die Maschine Daten analysiert, um zum Beispiel begleitend über einen Geschäftsprozess zu berichten, das Unterhalten mit einer Industriemaschine, mit einem Roboter, zu ermöglichen, mit smarten Geräten eine Konversation zu ermöglichen. Das ist ein Feld das wir gerade aufbauen. Da fällt einiges rein. Wir haben jetzt Prototypen aus dem Bereich connected-car: Meine E-Klasse kann jetzt mit mir schimpfen, wenn ich zu schnell oder aggressiv fahre. Aber nur, wenn sich das Ganze auch gegen meine Baseline verschiebt. Also wenn ich aggressiver fahre als sonst – das ist ein wichtiger Unterschied. Wir arbeiten auch gerade daran, mit einem größeren Chemiekonzern aus der Region Reportings der einzelnen Abteilungen diskutierbar zu machen. Dass sie also nicht nur das Scratch Sheet bekommen, die Visualisierungen dazu und die Linie. Stattdessen sollen sie dann eben auch zurückfragen können: Hey Reporting, wo sind deine Wendepunkte? Wie signifikant bist du im Vergleich zu anderen Abteilungen? Was macht der Trägermarkt? Und so weiter. Das sind alles Dinge, die Ihnen ein Graph so erst einmal nicht sagt, die Sie jetzt aber abfragen können. Und zwar in ihrer eigenen Reihenfolge, entlang ihres eigenen Erkenntnisprozess und eben auch so lange bis Sie es verstanden habe. Das macht im Moment ein bis drei Prozent unseres Umsatzes aus. Ist aber tatsächlich auch so ein Future-Business, das wir in den nächsten zwei Jahren komplett in einen Massenmarkt [aus]gerollt haben wollen.

Ab wann lohnt sich das Tool aus ihrer Sicht?

Da haben wir jetzt einen spannenden Punkt erreicht. Wir mussten das tatsächlich früher auf die Menge der Texte umrechnen. Da hatten wir immer die Faustregel 5000 und das Tool kostete 1000 Euro im Monat. Die Maschine lernt ja auch, die wird ja besser je länger man sie nutzt, d.h. man hat auch einen Anreiz, Texte wieder zu generieren. Dann hat man gesagt, 5000 Texte, dann kommen Sie auf 20 Cent pro Text, wenn Sie 1000 Euro im Monat bezahlen.

Jetzt haben wir das Packaging nochmal ganz drastisch verändert. Man kann die Maschine jetzt für 200 Euro im Monat pro Entwicklerplatz kaufen. Deshalb sagen wir jetzt: Sobald Sie den Verdacht haben, dass es ein Geschäftsmodell gibt, Sie wissen aber noch nicht wie sich das rechnet, Sie wissen den Business-Case dahinter noch nicht. Aber sobald Sie den Indikator haben, dass es einen Business-Case geben könnte, dann fangen Sie bitte an, diesen Seat zu buchen. Da kommt dann auch ein kleines Ausbildungsprogramm mit, in dem Sie dann auch lernen, wie Sie unsere Plattform und unsere Kommunikationssprache, das ATML3, korrekt bedienen. Anschließend können Sie damit einfach rumprobieren. Wenn eine gewisse Nutzungsmenge überstiegen wurde, kostet das nochmal 8 Cent pro Text. Damit ist der Preis jetzt proportional zur Textproduktion. So sind Sie extrem flexibel.

Sehen Sie das Tool eher für Redakteure oder für auch für Techniker?

Wir machen ja eigentlich zwei Sachen. Das eine ist: Wir konfrontieren Sie mit vielen Konzepten, die sich so ein bisschen an die Programmierwelt anlehnen. Das ist eine recht begrenzte Menge an Konzepten und wir bereiten diese so auf, dass Sie das als Journalist lernen können. D.h. also eigentlich passieren da zwei Dinge: Zum Einen bleiben sie bei dem Skillset, das Sie kennen: also Geschichten zu erzählen, Informationen zu transportieren, mit der Audience zu interagieren und gewinnen einen neuen Teil dazu, der sich eben auf AX Semantics bezieht.

Da bauen Sie einen generellen Skill auf, wie man eigentlich datengetrieben berichtet. Sie werden ja auch mit der Aufbereitung von Daten konfrontiert, wie man ein NLG-Problem erkennt, wo die Möglichkeiten von NLG eigentlich liegen und wie Sie die Strategien, die Sie in Zukunft für ihren Betrieb wählen, dann umdenken müssen, um die Vorteile von NLG ausnutzen zu können, um das in klingende Münze zu verwandeln.

Deshalb: Ja, es ist für Journalisten und es ist so gedacht, dass man ein bisschen interdisziplinären Business-Transfer hat, dass man eben ein paar IT-Konzepte lernt, aber dann eben auch einen nachhaltigen Skill hat, den man über AX Semantics hinaus auch weiter in seinem Haus hat.

Da gibt es auch Zertifikate dafür. Also wenn Sie das erfolgreich durchlitten haben, dann können Sie auch wirklich mittlerweile eine Prüfung machen, zumindest in den bezahlten Paketen. Da können Sie den Language Automation Basic, Text Automation Expert und Text Automation Professional auch in den Lebenslauf schreiben. Ich kenne mittlerweile auch Firmen, die verlangen das in ihren Ausschreibungen.

Wie groß schätzen Sie den Aufwand zur Einarbeitung?

Das kann ich Ihnen sogar relativ genau sagen. Die ersten Schritte machen sie üblicherweise nach zwei Tagen. Und bis sie das erste Projekt publizieren können, sehen wir einen Mittelwert der sich irgendwo unterhalb der Zwei-Wochen-Grenze bewegt. Bis sie sich also selbst on-geboarded haben und das erste Projekt zu einem Minimalwert gebracht haben, wo sie sagen, das ist jetzt auch für meine Nutzer einen Wert, da können Sie eher mit Wochen als mit Monaten rechnen.

Und danach gibt es aber einen Aufwand, den sie sich machen können, aber nicht müssen, um das Produkt dann kontinuierlich zu verbessern. Denn mit jeder Korrektur die Sie machen, lernt jeder Text, den Sie publiziert haben dazu, der kann dann nochmal gerendert werden mit dem Wissen, das die Maschine jetzt dazugewonnen hat.

Wenn jetzt ein Kunde sagt: Das Produkt möchte ich für meine Firma. Wie würden dann die weiteren Schritte aussehen?

Das ist erstaunlich unkompliziert geworden. Sie sollten jemanden benennen, der diesen Account bekommt und dann qualifiziert wird. Der kann sich mittlerweile Online anmelden. Müssen Sie einfach die Kreditkarte mitbringen oder eine Rechnung bei uns beantragen. Anschließend buchen Sie einfach ihren ersten Monat auf der AX Semantics Plattform für diese 199€. Buchen dann noch das On-Boarding-Packet dazu – das sind dann nochmal 2000€. Anschließend durchlaufen Sie unser eLearning und danach nochmal eine Schulung mit Entwicklern von uns.

Wenn Sie dann den ersten Kollegen qualifiziert haben, dann geht der mit allem was er jetzt weiß auf Themensuche und guckt sich an welche Anwendungsfälle es eigentlich gibt, welche Fachabteilung oder welches Ressort ihnen da gerne mal ein Thema lassen würde, sucht nach einer geeigneten Datenquelle und fängt einfach an. So machen Sie ihren internen Proof-of-Concept. Damit können Sie dann zum Vorstand oder zum entsprechenden Medienrat gehen und sagen: „Hey, ich habe da Folgendes durchgeführt, ich habe da folgende Erfahrungen mitgemacht, ich würde das gerne weiter ausrollen." Üblicherweise passiert dann, dass ein zweiter Kollege daneben sitzt und irgendwann auch ein dritter, die dann auch diese Qualifikation durchlaufen. Irgendwann bildet sich dann ein Nucleus heraus, der sich mit Textautomatisierung auskennt und der wandert dann auch von Thema zu Thema, von Projekt zu Projekt und reißt immer weitere Textprojekte an. Wenn dann mal ein ausreichend großes Team da ist, haben wir auch noch andere Schulungsformate, wo wir dann auch mal noch ganz komplizierte Sachen mit dem Team angehen können, wo wir dann komplexere Tool-Chains aufbauen können, wo die Kollegen dann zwei Tage lang mit uns einfach programmieren und alle vor Ort sind.

So was wie die Stuttgarter Zeitung zum Beispiel gemacht hat. Die haben jetzt Feinstaubberichte, die auf Opensource-Sensoren basieren. Das OK-Lab hat ein Bauplan für einen DIY-Feinstaubsensor mit Cloud-Anbindung veröffentlich. Diese Sensorwolke wurde an AX Semantics angebunden zusammen mit ein, zwei anderen Datenquellen und liefert jetzt für alle möglichen Locations in Stuttgart individuelle und zeitaktuelle Feinstaubberichte. Das sind dann die Projekte,

wo man dann sagt, da braucht man dann doch nochmal eine beratende Expertise von uns im Hintergrund. Die kriegt man dann auch, zum Beispiel in einer Pair-Programming Schulung über zwei Tage.

Dazwischen gibt es auch nochmal eine Sache. Wir haben mittlerweile eine relativ aktive Community. D.h. wenn Sie eine interessante Problemstellung haben, können Sie die auch einfach um die geschäftsrelevanten Bereiche bereinigen. Das müssen Sie eventuell ein bisschen schwärzen und anschließend können Sie das in unserem Nutzerforum posten. Teilweise helfen ihnen auch Nutzer aus anderen Firmen, die schon ähnliche Probleme hatten oder Nutzer von Unis. Da tummeln sich eigentlich alle mit ihren depersonalisierten Problemen drin. Das ist natürlich toll, weil so auch die Community was über Bedarfe und Problemstellungen lernt und Sie gleichzeitig schnelleren und ausführlicheren Support, als eben die auch ausführliche, aber eben limitierte Menge, die unser Supportteam mit seinen fünf Nasen leisten kann haben.

Und die NGL-Software funktioniert ausschließlich über die Cloud?

Ja, das ist im Moment noch so. Jetzt gerade auf Anforderung von der Versicherungswirtschaft hin bzw. auf Anforderungen aus dem IoT-Bereich hin, haben wir irgendwo so im nebligen Teil der Roadmap, also da wo man noch nicht so richtig auf Sicht fährt, da haben wir eigentlich so die Mission drin, dass wir gerne hätten, dass man unsere Plattform konfigurieren kann und dann einen Zustand herauspersistieren kann, also sprich alle Komponenten, die man braucht um das, was Sie trainiert haben, durchzuführen. Dass man die auch rausnehmen und in einen Worker packen kann, den man irgendwo lokal laufen lassen kann. Das ist aber ehrlicherweise noch nicht mal anentwickelt. Ich habe letztes Jahr gesagt, das sollte im Laufe von 2017 kommen. Jetzt können Sie sich vorstellen, was ich dieses Jahr sage: Es wird also wahrscheinlich im Laufe von 2018 kommen.

Wie sieht es dann mit dem Datenschutz aus? Wenn ich Daten hochlade, wo landen die dann?

Wir können das sogar relativ genau sagen. Wir laufen in der Amazon AWS und haben einen Vertrag mit Amazon gemacht, genau aus dem Grund. Die garantiert uns, dass wir in Frankfurt am Main laufen. Wenn wir eine besondere Privacy-Anfrage für einen Kunden schalten, laufen wir nur in Frankfurt am Main, ansonsten auch nochmal in Rotterdam. Aber auf jeden Fall innerhalb der Safe-Harbor-Zone.

Dann kommt noch dazu, dass wir Komponenten gewählt haben, die HIPAA Compliance sind. Das ist ein Datenschutzstandard für amerikanische Healthcare-Anwendungen. Das bedeutet dass, wenn Sie in Deutschland eine medizinische Anwendung machen, dann braucht es ein Einzelfallgutachten des Landesdatenschutzbeauftragten des Bundeslandes, in dem Sie diese Anwendung zum Einsatz bringen, sowie jedes einzelne Krankenhauses, indem das Ding dann läuft. Der freut sich natürlich auch, wenn er darauf deuten kann, dass die Komponenten irgendwann schon mal für irgendwas zertifiziert wurden. Das ist in unserem Fall, da es ein amerikanischer Cloud-Provider ist, eben der amerikanische Standard HIPAA.

Ansonsten gibt es aber eine ganz einfache Praxis, gerade wenn Sie es mit personenbezogenen Daten zu tun haben. Im Zweifelsfall können wir einen ähnlich guten Text schreiben, ob wir jetzt wissen, ob der Herr, über den wir schreiben, ein Herr oder ein Herr Kunze ist. Und dann sagt unsere Plattform: Lieber Herr #251195F69 und ihre Plattform, die den Text dann ausspielt, fügt an dieser Stelle dann zum Beispiel den Klarnamen ein.

Warum gibt es für die NLG-Plattform die eigene Programmiersprache ATML3?

Wenn Sie versuchen, das alles in Python abzubilden, dann müssen Sie ja eigentlich so Dinge, wie zum Beispiel die Morphologie in einer Sprache, die grammatikalischen Zusammenhänge, die wir jetzt im Hintergrund realisieren, wenn Sie so ein ATML-Objekt entwerfen, oder die verschiedenen Auswertungstypen die Sie machen, müssten Sie ja alle jedes Mal neu implementieren, oder Sie müssten sich selbst eine Software schreiben, die sie implementieren kann und sie dann auch wieder konfigurieren. Bei uns war es das Gleiche. Wir haben auch eine Software, die ist mittlerweile in Python. Die war aber früher in Visual Basic. Wir haben die am Anfang auch jedes Mal neu gestrickt. Um jeden Text herum haben wir ein Skript geschrieben. Aber dann haben wir gesagt: Das sind doch alles Regelmäßigkeiten, das geht doch einfacher. Ich will nur noch sagen müssen, dass der Fernseher und nicht Bildschirm sagen soll. Aber ich muss nicht jedes Mal entscheiden: „if Dativ, dann dem Fernseher“ oder „if Genitiv, dann des Fernsehers“.

Das ist auch so eine Skalierungssache. Wir abstrahieren die Programmiersprache aber immer weiter vom Nutzer weg. Also an vielen Stellen sieht man nur noch den Code, der geschrieben worden wäre, aber man schreibt ihn nicht mehr selber. Gerade wenn man ein Wort einbindet und die grammatischen Parameter setzt. Das ist schon komplett wegabstrahiert. Das wird in der nächsten Generation noch ein bisschen vertieft. Wir haben jetzt gerade ein neuronales Netz laufen im Prototypenzustand, das ihre Eingabesätze, die Sie machen auf Dependenzen zerlegt, komplett durch annotiert und alle grammatischen Zusammenhänge und Erbschaftsbeziehungen in diesen Sätzen auch schon annotiert. Sodass sie nur noch Bezüge zu bestimmten Daten oder Vokabeln, die Sie im Hintergrund gemacht haben, also Gruppe, Property, Lookup, Textcontainer, festlegen müssen. Das können sie dann darauf ziehen. Das ist ein Snap-In-System. Aber Sie müssen eben nichts mehr schreiben an dieser Stelle bzw. konfigurieren. Wir sind jetzt zwei Metaebenen drüber. Das neuronale Netzwerk analysiert ihre natürliche Sprache und sagt voraus, wie die Regeln zur Generierung von natürlicher Sprache in Maschinensprache, nämlich in ATML3 aussehen müssten.

Das ist jetzt Namedropping: Das ist eine hybride KI-Lösung, wo ein Teil der Value-Chain durch echte, tief lernende und redaktionelle KI gelöst wird. Aber der Teil, der zur Laufzeit ihre Texte produziert, solide Automatisierung, ist ein festes Regelwerk. Warum ist das wichtig? Das ist deshalb wichtig, weil vollredaktionelle Systeme sich ihrer Verantwortung entziehen.

Wenn Sie etwas haben, das auf jeden Input trainiert wird, dann kann das System getrollt werden, Sie sind aber rechtlich nach wie vor dafür verantwortlich, dass Sie es ja in den Verkehr bringen und rausgeben. Da sind meine geschätzten Freunde aus dem Microsoft Developer-Center in

München mal letztens in so eine Sache hineingerannt. Sie erinnern sich vielleicht an die Chatsoftware „Tay“. Ein Bot in Gestalt einer jungen Dame, der aus den Anfragen, die er auf Twitter bekommen sollte, lernen sollte, wie man die richtig beantwortet. Twitter besteht leider immer noch zu 60% aus Trollen. Und am ersten Tag ist Tay vom Netz genommen worden, als er sagte: „I believe essentially Hitler was right!“. Das war das erste was er sagte, bzw. das letzte was er sagte, bevor er das erste Mal vom Netz ging. Dann gab es eine riesige Wartungsschleife. Tay wurde mit allen möglichen Filtern ausgestattet und sagte dann nach einer glorreichen Woche in Betrieb: „I believe all feminists should burn in hell“. Dann ist er nochmal eine Woche vom Netz genommen worden und mit einer Reihe anderer Filter, teilweise redaktionellen und neuronalen Netzen basierten Filtern ausgestattet worden. Und dann ging er endgültig vom Netz mit den Worten: „I smoke weed in front of the police cops!“. Natürlich kann man das jetzt lesen, wie: Der regelbasierte Programmierer zieht alle Schreckensbeispiele heraus, um das Konkurrenzparadigma zu diffamieren, so wie es Tesla und Edison auch gemacht haben mit ihrem Stromkrieg. Worauf ich eigentlich hinaus will ist, dass Sie einen Teil brauchen der persistent ist, damit Sie einen Zustand haben, mit dem Sie irgendwo transparent machen, für den Sie Verantwortung übernehmen oder den Sie auch im Zweifelsfall irgendwo einreichen und hinterlegen können.

Und diesen Zustand so trivial wie möglich herzustellen, das Regelwerk so einfach wie möglich zu erstellen, das lösen wir jetzt verstärkt in der nächsten Generation, die März/April rausrollen wird. Dann fast nur noch mit Hilfe von Sprachanalyse auf Basis von Google SyntaxNet, auf Basis von neuronalen Netzen.

Aber letztendlich hat immer noch der Anwender den Algorithmus im Griff?

Genau. Der Algorithmus manifestiert sich am Ende in einem Regelwerk in ATML3. Also wenn man ATML3 lesen kann, weiß man ziemlich genau, was die Maschine gerade denkt. Hinzu kommt, dass so ein Regelwerk debugt werden kann. Also man kann sich anschauen, an welcher Stelle vom Entscheidungsbaum eigentlich gerade welcher Wert rauskommt. Es gibt den ATML3-Debug-Tree. Für jeden generierten Text kannst du eigentlich sehen, was genau die Maschine an welcher Stelle darüber gedacht hat, also welche deiner Beobachtungen eigentlich wahr wurde, und was weitergegeben wurde. Das werden wir auch noch graphisch aufbereiten. Wir nennen das Projekt im Moment „Spagetti-Monster“, weil es erstmal ganz viele Linien darstellt. Aber dann kannst du auch sehen: Mein Datum geht von hier in diese Auswertung, da kommt folgende Liste raus, aus der wird Folgendes gefiltert und das kommt im Text an. Da hast du dann einen Einblick, den du in einem reinen neuronalen Netzsystem überhaupt nicht haben kannst. Das ist stattdessen dann die Magic Blackbox. […]

Thema: Sprachsteuerung

Viele schlagen sich damit rum, dass man jetzt auch kontextuelle Medien anbieten sollte, weil es sozusagen eine junge Kohorte nachschiebt, die es nur noch gewohnt ist, mit Bots zu sprechen. Auch das können wir mittlerweile. Wir können auch Alexa und GoogleHome-Anwendungen. Also auch den Sprachskill, der dich zum Beispiel zu einem Inhalt navigiert oder eine etwas freiere

Art erlaubt nach dem Wetter oder Servicecontent zu fragen. Auch das ist etwas, was wir durchdenken können. Die sind nicht ganz billig. Mittlerweile sitzen da noch die teuren Superhigh-Potentials 30 Tage dran. Aber die sind dann auch richtig gut. Aber das wäre auch nochmal eine Möglichkeit ein Innovationssignal zu setzen. Dass man sagt, man hat jetzt nochmal ein ganz neues Medium angerissen. Da entwickeln wir auch gern dran mit.

Wie kann ich mir das vorstellen?

Du kennst doch den typischen Alexa-Skill. Du musst relativ spezifisch fragen: „Alexa, wie wird heute das Wetter in Stuttgart?“ und dann gibt dir die Alexa normalerweise eine super-generische Antwort und wirft mit allem, was sie weiß. Wir haben das Paradigma umgekehrt und Alexa die Zunge rausgeschnitten und haben gesagt: Da wo sie erkannt hat, was du eigentlich von ihr willst, darf sie nichts mehr sagen, sondern gibt ihre Daten über API an unsere Plattform weiter und die darf dann nochmal nachdenken bevor sie antwortet, was Alexa nicht tut. Sprich: Alexa fischt nur eine Antwort aus einem Bucket, füllt einige Slots drin und gibt das dackeldoofste zurück, was sie halt sagen kann.

Unsere Anwendung versucht Schlüsse zu ziehen, was es dir wiederum erlaubt, viel freier zu fragen. D.h. du musst nicht mehr fragen: „AX wie wird das Wetter heute in Stuttgart?“ sondern du kannst fragen „Kann ich heute was auf den Grill werfen?“ Wenn wir irgendwas nicht richtig verstanden haben, fragen wir nochmal nach: „Wo möchtest du denn grillen?“ und anschließend kriegst du eine Antwort: „Ja, du kannst heute grillen. Aber es könnte ein bisschen kalt sein!“

4. Technische Daten: Mevo-Kamera (Auszug)[145]

Dimensions & Weight

Höhe	63 mm / 2.5 inches
Breite	50 mm / 2 inches
Gewicht	142 g / 5 ounces
Farbe	Black
Interfaces	micro USB 2.0, micro SD slot

Video

Sensor	Sony® 4K Sensor
Processor & Encoder (SoC)	Ambarella A9SE
Sensor Resolution	12.4 Megapixels
Capture Resolution	4K - 3840 x 2160 pixels
Aspect Ratio	16:9
Frame Rate	30fps
Video Codec	H.264

Built-in Microphone

Microphone Type	Dual analog MEMS, high signal-to-noise ratio (65 dB), matched
Codec	8kHz -192kHz, 8/10/12 bits

Audio

Audio Mixer	Stereo/Mono, Meter, Gain control, 2 inputs - internal stereo microphones and external audio input
Audio codec	AAC

Networking

Wi-Fi	802.11ac @ 2.4Ghz / 5GHz. 2x2 MIMO, dual antennas Access point mode or join an existing Wi-fi network
Wi-Fi Network Required	No, the camera can generate it's own network (access point) and stream via the mobile LTE connection
Bluetooth	Bluetooth 4.1 - BT and BLE compatible
Ethernet	100 Mbps ethernet – Requires Mevo Boost accessory

[145] Vgl. https://getmevo.com/tech-specs/camera (Stand: 22.12.2017).

5. Interview Aeaneas Rooch (Scisys)

Das Interview mit Aeneas Rooch von Scisys fand am 03.01.2018 von 9.00 Uhr bis 9.20 Uhr via Telefon statt. Aeneas Rooch ist Mathematiker und arbeitet bei Scisys Deutschland GmbH als Solution Manager.

Was war der Auslöser diese Software zu entwickeln?

Mein Job besteht unter anderem darin, sich Gedanken zu machen, was wäre denn mal wichtig zu haben. Ich arbeite in der Abteilung „Forschung und Entwicklung" und bin im Wesentlichen für beides zuständig. Mehr mit Fokus auf Forschung, dass ich Sachen prototypisch ausprobiere und evaluiere. Mein Job ist es, sich Gedanken zu machen: Was könnte denn mal eine schöne Funktion sein? Was könnte man denn mal bauen, was in fünf Jahren interessant sein könnte? Ich beobachte also, was in den Medien passiert und was in der IT-Branche passiert.

Ein großes Thema seit vielen Jahren ist künstliche Intelligenz. Wir alle haben künstliche Intelligenz auf unseren Smartphones und die meisten nutzen das auch auf irgendeinem Wege. Ob das jetzt darum geht, dass die Worterkennung beim Tippen von Nachrichten einem vorschlägt, was man gerade schreiben will. Da braucht man nur ein zwei, drei Buchstaben eintippen und in der Regel kommen dann die passenden Wörter. Da hat das System von anderen Nutzern schon gelernt, was das nächste Wort sein könnte. Bis hin zur Spracherkennung, d.h. wenn wir ins Telefon sprechen und sagen: „Hey Siri, bitte beim SWR anrufen“ oder „Wie ist das Wetter morgen in Stuttgart?“. Dann funktioniert die Spracherkennung ja auch nicht nach festen Regeln, sondern dahinter steckt maschinelles Lernen. Das sind Technologien, die nutzen wir ganz selbstverständlich. Ich habe mir überlegt, wo kann das hilfreich sein und zwar ganz konkret im Radio? Viele Forschungsprojekte haben einen akademischen Charme, dass man so sagt: „Das könnte man mal machen“ und in zehn Jahren ist es vielleicht soweit. Aber wo könnte man es wirklich brauchen?

Ich hatte mir überlegt: eine Arbeit, die wir immer gern Computern überlassen ist etwas, das monoton und langwierig ist, da geben wir gerne Kontrolle ab. Und da, wo wir gerne Kontrolle behalten, ist bei kreativen Prozessen. Wo wir sagen: Das möchte ich selber gestalten. Ich hatte mir überlegt, das könnte spannend sein beim Audioschneiden. Das tägliche Geschäft des Redakteurs oder Radioautors ist es ja, Aufnahmen die man selbst gemacht hat, oder draußen oder im Studio, schön zu schneiden. Im einfachsten Fall ist das: Ich habe selber was im Studio eingesprochen und muss das Papierrascheln am Anfang wegschneiden, oder wenn ich mich versprochen habe. Im komplizierteren Fall ist das: ich habe draußen vor Ort bei einer Aktion etwas aufgenommen und dann möchte ich daraus nur zehn Sekunden für einen Beitrag verwende. In ganz vielen Fällen sind das Standardsachen. D.h. die Fälle, wo man wirklich diffizil schneidet, wo man wirklich ins Detail gehen muss, geschickt blenden muss, die sind eher die seltenen Fälle. Die meisten sind eher die: Da hat jemand sich versprochen und fängt vorne nochmal an oder ich muss einen Räusperer rausschneiden. D.h. das sind so Standardfälle.

Da habe ich mir gedacht: Kann da nicht eine Maschine lernen, was da die perfekte Schnittposition ist? Und der erste Fall, den ich mir überlegt habe – man muss ja langsam anfangen und man darf nicht gleich zu hoch greifen – ist: Der Mensch sagt: „Hier ungefähr, bei der Position, da ungefähr möchte ich schneiden, da hat der Mensch sich versprochen, da möchte ich einen Schnitt setzen." Die Maschine findet dann die optimale Position in der Nähe.

Ich muss schon ungefähr vorgeben: Von da bis da möchte ich ungefähr den O-Ton haben. Aber den Feinschnitt den hoffe ich, kann dann ein Algorithmus und zwar anhand von bereits gesetzten Schnitten. Ich möchte also, dass ein Algorithmus auswertet, wie Menschen Schnitte setzen und daraus selber lernt, wo eigentlich eine gute Schnittposition ist.

Und wie schwer war es, einen geeigneten Algorithmus zu finden?

Das weiß ich noch nicht. Das ist wie immer in der Wissenschaft. Es gibt viele Ansätze. Ich bin kein Experte im Bereich der künstlichen Intelligenz. Ich weiß gar nicht genau, was es da für Möglichkeiten gibt. Aber ich weiß, es gibt viele verschiedene Herangehensweisen, wie man so ein Modell aufstellt und wie man so eine künstliche Intelligenz trainiert. Es geht ja darum, dass man generell bei künstlicher Intelligenz, einem Algorithmus bestimmte Daten vorsetzt anhand derer der Algorithmus lernen kann. Man sagt also: Hier das ist gut, das ist schlecht! Und nach einer gewissen Zeit hat man den Algorithmus trainiert. Danach gibt man ihm Daten, die nicht mehr die Klassifikation verraten.

Es geht im Wesentlichen immer ums Klassifizieren. Ich lege dem Algorithmus etwas hin und sage: Jetzt sortiere du mir das mal in die richtige Schublade ein. Im einfachsten Fall in die Schublade „gut" oder „schlecht". Das geht natürlich beliebig kompliziert. Zum Beispiel bei der Spracherkennung in die Schublade: Welches Wort ist das? Welcher Satz ist das? Welcher Laut ist das? Man trainiert, indem man Datenbeispiele hinlegt und dazu verrät, wie man die einsortieren muss, ohne dass man verrät warum. Wenn man wüsste warum, wenn man die genauen Regeln formulieren könnte, wäre es immer einfacher, die Regel exakt zu implementieren. Es geht ja darum, dass man sagt: Die Regel ist vielleicht so kompliziert oder vielschichtig, dass ich die gar nicht formulieren kann. Irgendwann, wenn man das anhand von einem großen Datensatz von Trainingsdaten, so heißt das, gemacht hat, dann hofft man, dass man dem Algorithmus auch etwas vorsetzen kann und er macht die Klassifikation alleine anhand der Dinge, die er vorher gelernt hat.

Das ist die erste Phase des Projektes, oder vielleicht schon die zweite, wenn man kleinschrittig denkt. Die erste ist erstmal Daten sammeln und die zweite ist: Verschiedene Ansätze auszuprobieren um zu gucken: Kommt da überhaupt was bei raus? Selbst wenn man es gut macht, kann man das Pech haben, dass man ganz viele Daten sammelt und den Algorithmus lernen lässt, aber trotzdem kriegt er das am Ende nicht hin. Man kann es auch schlecht machen. Dann hat man Daten gesammelt und bekommt Unsinn heraus. Aber selbst, wenn man es gut macht, ist es Ausprobieren.

Und wie ist gerade der aktuelle Stand? In welcher Phase befinden Sie sich gerade?

Mitten in der ersten Phase, wenn wir kleinphasig denken. Oder am Anfang der ersten, wenn wir großschrittig denken. Das Datensammeln ist tatsächlich ein großes Problem. Das hätte ich mir nicht so vorgestellt, das ist eine Herausforderung. Und zwar nicht, was ich eigentlich gedacht hätte, datenschutzrechtlicher Natur. Ich hatte eher befürchtet, die Sender haben Angst um ihre Daten. Aber tatsächlich habe ich einen Sender gefunden, mit dem wir eine Datenschutzvereinbarung unterzeichnet haben und ein Non-Disclosure Agreement. Das Deutschlandradio will prototypisch evaluationsmäßig bei uns mitmachen und mit uns forschen. Die setzen auch unsere eigene Software ein.

Aber es ist relativ schwierig, Daten zu finden, an denen ein Algorithmus, wie auch immer er nachher aussehen mag, lernen kann. Also ich habe mir überlegt: Wir haben in unserer Software auch ein Schnittprogramm. Da setzen Redakteure, Mediengestalter oder Tontechniker jeden Tag hunderte bis tausende von Schnitten, vielleicht sogar mehr. Aber wie komme ich überhaupt da dran? Woher weiß ich, wo die Schnitte gesetzt worden sind? Denn wenn das Projekt fertig ist, wenn man seinen Audiobeitrag fertig geschnitten hat, dann hat man diese Daten nicht mehr. Dann rendert man ein fertiges Audio. Man führt alles zusammen und generiert ein neues Audiofile in dem keine Informationen mehr über die alten Versatzstücke und Schnitte drin sind. Deswegen musste ich mir Gedanken darüber machen: Wie komme ich denn überhaupt an diese Schnitte?

Eigentlich ist es einfach, weil die Schnittsoftware ja uns gehört. Aber die Herausforderung war, diese Schnitte zu bekommen, ohne die Schnittsoftware anzufassen. Bei großen Kunden im Radiogeschäft, die tagesaktuell arbeiten, ist es immer eine heikle Sache, so essentielle Komponenten wie den Schnitteditor zu ändern. D.h. Ich hätte natürlich den Editor ändern können, dann hätte man in sehr großem Maß testen müssen, ob alle bisherigen Funktionen nach wie vor zur Verfügungen stehen. Dann hätte der Kunde das entsprechend lange getestet. Das hätte sehr viel Zeit gebraucht. Die Herausforderung war also die Daten abzugreifen ohne das Schnittprogramm zu ändern. Das haben wir jetzt gemacht. Die Software liegt gut getestet bei uns und wartet darauf, dass wir sie in den nächsten Tagen dann tatsächlich beim Kunden installieren. Ich hoffe, dass wir in einem Monat tatsächlich soweit durch sind, dass wir in den Regelbetrieb gehen. Dass wir also ab Ende Januar fortlaufend einen kontinuierlichen Datenstream bekommen. Also der kurzfristige Plan für die nächsten Tage dieser Woche ist: letzter Feinschliff bei uns. Nächste Woche dann Test beim Kunden. Die wollen die Software natürlich auch testen, bevor sie sie auf den ihre ganzen Rechnern installieren.

Und dann heißt erstmal: Daten sammeln und die Maschine lernen lassen?

Genau. Dann heißt es erstmal Daten sammeln. Dann müssen wir uns noch Gedanken machen, wie wir die Daten, die wir gesammelt haben eigentlich zu uns und den beteiligten Wissenschaftlern kriegen. Da gibt es ja auch mehrere Möglichkeiten. Da will ich genau wissen, wie viele Daten dabei rumkommen. Die Schätzungen reichen von 10 GB bis 100 GB am Tag. Das wissen wir ein-

fach nicht. Je nachdem wie groß das ist, gibt es verschiedene Möglichkeiten. Entweder wir übertragen das per FTP oder wir laden das auf einen Cloudspeicher hoch. Eine andere Möglichkeit wäre, dass wir das per Hardware machen. Das klingt jetzt etwas altbacken. Aber selbst bei AWS, bei Amazon, wenn Sie da ihr Datencenter in die Cloud laden wollen und Sie sagen: ich habe hier so und so viel Tera- oder Petabyte Daten, dann verlangt keiner, dass Sie das per Internet hochladen. Dann schickt ihnen Amazon auch eine Festplatte, um das da drauf zu kopieren. Die Möglichkeit müssen wir noch evaluieren und dann müssen wir abwarten. Erstens wie viel rumkommt und zweitens wieviel Daten wir brauchen, bis die Wissenschaftler sagen können: „Das ist gut und das ist schlecht."

Danach ist es ein iteratives Verfahren. Im Idealfall lassen wir einen Monat lang die Maschine laufen und danach sagen die Datenwissenschaftler: „Perfekt, das kann so weiterlaufen" und wir haben auch schon was gelernt. Im schlimmsten Fall sagen die: „Wir kriegen da überhaupt nichts raus. Das bringt überhaupt nichts." Und dann müssen wir herausfinden warum? Ob das daran liegt, dass vielleicht der Ansatz des maschinellen Lernens schlecht ist oder ob die Struktur der Daten das nicht hergibt.

Es steht und fällt natürlich alles mit der Qualität der Daten. Man sagt, Daten müssen sehr gut strukturiert sein. Wenn Sie genau wissen, was für Daten Sie haben und wie die aufbereitet sind, dann ist das schon mal gut. Das kann ich garantieren, weil das ja unsere Daten sind. Ich habe hier nicht irgendeinen unbekannten Datensatz und muss rumraten, sondern ich kann die so aufbereiten, wie die Wissenschaftler das haben wollen. Ich kann also sagen, dass das, was die Wissenschaftler unwichtig finden, was vielleicht das Lernen behindert, das filtere ich gleich von Anfang an schon raus. Ich kann aber auch sagen: „Wir lassen das und das noch drin, obwohl wir das unwichtig finden." Das können wir sozusagen designen. Aber es kann trotzdem sein, dass die Informationstiefe nicht ausreicht. Denn zum Beispiel können wir Mausbewegungen im Augenblick nicht nachverfolgen. Wir können also gar nicht sehen, wo jemand den Schnitt setzt und vielleicht nochmal umsetzt. Diesen Prozess sehen wir nicht.

Das, was wir sehen können, was wir an Daten rauskriegen ist: wo wurde der Schnitt gesetzt und durchgeführt? Wenn jemand einen Schnitt setzt und rückgängig macht, dann kriegen wir das auch nicht raus. Wir kriegen in der Regel nur die gesetzten Schnitte raus. Es mag sein, dass das ausreicht – das will ich sehr hoffen – um zu lernen, wo ein guter Schnitt ist. Der Nachteil bei dieser ganzen Sache ist: Wir kriegen keine Information darüber, wo eigentlich ein Schnitt schlecht ist. D.h. weder sehen wir was der, der vor dem Rechner sitzt, rückgängig macht noch bekommen wir eine Bewertung. D.h. bei einem echten Trainingssatz sagt man normalerweise: Hier ist ein Beispiel und das war gut, hier ist ein Beispiel das schlecht ist, usw. Oder wenn ein System Handschriften lernt, kann man auch sagen: Das ist ein „A", das habe ich als Mensch erkannt, das ist ein „E", das ist ein „I", das nächste weiß ich nicht, das ist gekritzelt. Diese Informationen fehlen uns. Alles, was wir sagen können ist: Hier ist ein geschnittenes Audio und da waren die Schnitte und wir müssen implizit annehmen, dass alle gesetzten Schnitte gut sind. Wenn da jetzt irgendein Trottel sitzt und total lieblos schneidet, oder vielleicht geht es auch nicht anders, ich will ihm da

nichts Böses unterstellen, dann haben wir halt einen Datensatz da drin, der nicht schön geschnitten ist. Wir können leider nicht sagen: Und das hier war ein schlechter Schnitt!

Das wir dann im Endeffekt auch als guter Schnitt gewertet.

Genau. Also wir haben den Sonderfall, dass alles was wir als Trainingsdaten haben, Positivbeispiele sein müssen. Ein Mensch könnte sich das ganze Material anhören und dann nochmal bewertet. D.h. wir müssten das aufbereiten. Wir kriegen an Daten die Originalaudios und alle gesetzten Schnitte. Aus den Daten kann man das finale Audio rekonstruieren. D.h. was wir machen könnten ist, wir könnten das finale Audio neu rekonstruieren lassen, wir könnten alle Schnitte anzeigen und dann müsste sich das jemand anhören und optisch anschauen „Ah, da war ein Schnitt" und dann bewerten „Habe ich den gehört oder habe ich den nicht gehört."

Aber könnte man nicht auch einfach willkürlich irgendwelche Schnitte setzen, die wirklich schlechte sind, und die dann als „schlechte" Schnitte bewerten lassen?

Oh, das ist eine sehr schöne Idee. Ja, das könnte man tatsächlich. Das müsste ich mal den Wissenschaftlern vorschlagen, was die dazu zu sagen. Dass man künstlich noch schlechte Informationen mitgibt und sagt: Hier, das ist was Schlechtes!

Da muss ja dann auch niemand mehr drüber hören. Das wird dann richtig schlecht sein.

Ja, das ist dann denkbar schlecht. Einfache Zufallsschnitte. Dann hätten wir negativ Beispiele. Das ist eine gute Idee. Die nehme ich gerne mit. Das müssen dann tatsächlich die Datenwissenschaftler beurteilen, weil die mehr Erfahrung haben. Das ist für alle ein bisschen ausprobieren. Ich kenne mich mit den Modellen nicht aus. Es kann sein, dass es Modelle gibt, die sagen: Nein, wir sind darauf spezialisiert nur aus Positivbeispielen zu lernen. Ich weiß ja nicht wie menschliches Lernen genau funktioniert. Es gibt ja beides. Im Unterricht sagt der Lehrer: So geht es. Der Lehrer zeigt Positivbeispiele. Man fängt immer damit an, dass man Positivbeispiele gibt. Erst durch das Selbermachen, wenn man selber Fehler macht und der Lehrer sagt: „Nein, das macht man nicht so, das geht anders." Nur durch so eine Korrektur lernt der Mensch. Aber man fängt trotzdem immer mit den Positivbeispielen an. Also es mag sein, dass es da Lernmodelle für maschinelles Lernen gibt, die keine Negativbeispiele brauchen. Das weiß ich aber nicht. Das gebe ich aber gerne weiter. Schöne Idee.

Bei unserem Telefonat im Dezember haben Sie gesagt, dass Interviews dadurch geschnitten werden, dass sie in ein Schreibprogramm wie Word eingefügt werden und anhand der Transkription geschnitten werden.

Da würden wir zwei Dinge kombinieren. Das eine ist dann tatsächlich das, worüber wir gerade gesprochen haben. Ich nenne das mal die „Feinpositionierung von Schnitten". Das andere ist klassisches Speech-to-Text. Das ist ein ganz klassischer Anwendungsfall von künstlicher Intelligenz. Das gibt es auch schon: Wenn Sie bei Siri was diktieren, können Sie einen Service im Internet buchen, dass Audios in Text verwandelt werden. Nehmen wir mal an, das funktioniert gut, ist

meine Vision, wie man das mit einem, ich nenne das mal „Schnittfix“ kombinieren kann: Sie haben ein Interview, das haben Sie aufgenommen und dann lassen Sie sich das transkribieren. Dann bekommen Sie das Transkript auf den Bildschirm. Jetzt wollen Sie einen bestimmten Satz rausschneiden und erkennen im Transkript: Von da bis da, das sind die Wörter von dem Satz, der mich interessiert! Wenn Sie auf ein bestimmtes Wort suchen, dann wissen Sie ja über das Transkript, welchem Zeitcode das im Audio entspricht. D.h. So können wir in Word einen Satz markieren.

Dadurch dass Sie die Wörter markiert haben, wissen Sie von wo bis wo das ungefähr in ihrer Audiodatei ist, Sie wissen also Startpunkt und Endpunkt. Das ist ja nichts anderes, als hätte ich manuell einen Start- und Endpunkt vorgegeben. Wenn Sie jetzt diesen Schnittfix los lassen, dann sagt der: ich erkenne den Startpunkt und ich erkenne den Endpunkt. Jetzt mache ich die Feinjustierung. Der perfekte Schnitt für diesen Startpunkt und Endpunkte wäre plus minus so und so viele Millisekunden nach vorne und nach hinten. Dann bekommen Sie ein sehr sauberes Audio, aus dem nur der ausgewählte Satz aus dem Transkript rausgeschnitten wurde. Das würde komplett überflüssig machen, dass wir mit der Hüllkurve arbeiten. Zumindest in einfachen Fällen.

Wir können uns schon Situationen vorstellen, in denen das überhaupt nicht klappt: Ein Text mit vielen Fremdwörtern, wo die Worterkennung schlecht ist, ein Audio bei dem die Leute durcheinanderreden oder ein Audio mit Hintergrundgeräuschen. Überall da, wo das Transkript schlecht ist, wo diese Einparkhilfe nicht funktioniert, da wird das fehlschlagen. Aber in einer sauberen Aufnahme, in einem Raum ohne Hintergrundgeräusche, in einem Studio, da kriegen Sie ein sehr gutes Transkript. Da können Sie gut in dem Text arbeiten. Dann brauchen Sie sich nicht mehr durch die Hüllkurve angeln. Sie können mit einem Blick am Text erfassen, worum es eigentlich geht und können dann wie in Word einen Satz auswählen und den Satz ausschneiden.

Das ist ja kein Kunststück, dass Sie sagen: Ich habe Speech-To-Text und ich habe ein Audio. Jetzt markiere ich wie in Word einen Satz, und den möchte ich gerne rausgeschnitten haben. Man kann das sozusagen matchen: Das Wort, das an dieser Stelle erkannt worden ist und das andere Wort ist bei Sekunde so und so erkannt worden. Dann kann man von da bis da das Wort herausschneiden. Nur bekommen Sie dann keinen schönen Schnitt, weil genau auf der Millisekunde, auf der ein bestimmtes Wort erkannt wurde, mit Sicherheit keine schöne Position zum Schneiden ist. Das wird dann mitten im Wort sein oder sogar Mitten im Laut.

Haben Sie noch Anmerkungen?

Das ganze Projekt hat nur Erfolg, je mehr Daten wir haben. Das ist sozusagen die größte Arbeit im Vorfeld: Daten sammeln. Leute finden, die da mitmachen wollen. Ich hoffe sehr, dass da noch andere Anstalten folgen. Ich freue mich sehr, dass das Deutschlandradio mitmacht. Aber wir haben ja noch andere Kunden, es gibt ja noch ein paar mehr Sender in Deutschland. Dass auch alle die erkennen, dass sich das lohnt da mitzumachen.

Ein Punkt vielleicht noch, der vielleicht wichtig ist: Wenn wir die Daten sammeln, dann werden die natürlich anonymisiert. Da sitzt zwar ein Mensch vor dem Computer und schneidet. Aber wir speichern nicht, wie die Audiodatei heißt, wir speichern nicht wie der Arbeitsplatz heißt, unter welchem Benutzernamen der eingeloggt ist. Alles, was wir speichern, ist das Originalaudio und die Position der Schnitte. Sie haben ja bei Audio, nicht so wie bei einem normalen User, der im Privatbereich ein Audio hat, nur eine Datei mit einem Dateinamen, sondern im Hörfunkbereich haben Sie ja eine Audiodatei plus eine umfangreichen Metadatensatz, wo dann drinsteht: Wer spricht hier? Worum geht es? Für welche Redaktion ist das? Wann wurde das erstellt? usw. All diese ganzen Metadaten werden gelöscht, weil wir das möglichst anonym haben wollen. Das einzige, was übrigbleibt, das Originalaudio, damit kann man sich immer noch anhören kann. Das können wir nicht verschlüsseln, denn daraus muss man lernen, und die Schnittposition. Möglichst viel schmeißen wir weg, um den Leuten das gute Gefühl auch zu geben, dass da nicht irgendwie Datenklau betrieben wird, dass da nicht irgendwie der gläserne Redakteur Angst haben muss vom Sender sekundengenau überwacht wird.

6. IBM Bluemix: Cloud-Services (Auszug)

- **Watson Conversation**: Build and deploy chatbots and virtual agents
- **Watson Discovery:** Uncover connections in data by combining automated ingestion with advanced AI functions
- **Speech-to-Text:** Easily convert audio and voice into written text
- **Text-to-Speech**: Convert written text into natural-sounding audio in a variety of languages and voices
- **Language Translator:** Dynamically translate news, patents or conversational documents
- **Natural Language Classifier:** Interpret and classify natural language with confidence
- **Natural Language Understanding:** Analyze text to extract metadata from content such as concepts, entities and sentiment
- **Visual Recognition:** Tag, classify and search visual content using machine learning
- **Tone Analyser:** Analyze emotions and tones in written content
- **Personality Insights:** Predict personality characteristics, needs and values through written text[146]

[146] Vgl.: IBM Cloud Products (Quelle: https://www.ibm.com/cloud/products/#ai (Stand: 10.12.2017)

7. Ergebnisse IBM Speech-to-Text

Test 1: Sprecher

Our world is changing in many ways and one of the things which is gonna have a huge impact on our future is artificial intelligence AI. Bringing another industrial revolution. Previous ~~industrialization~~ expended humans mechanical power. This new revolution this second machine age. Is going to expand our cognitive abilities are mental power. Computers ~~are~~ not just going to replace manual labor but also mental labor.

Test 2: Sprecher mit Straßenatmo

Our world is changing in many ways and one of the things which is gonna have a huge impact on our future is artificial intelligence AI. Bringing another industrial revolution. Previous ~~doctrine~~ revolutions expended humans mechanical power. This new revolution this second machine age. Is going to expand our ~~call me to. The teens are~~ mental power. Computers are not just going to replace manual labor but also ~~mentally~~.

Test 3: Sprecher mit weißem Rauschen (leise)

Our world is changing in many ways and one of the things which is gonna have a huge impact on our future is artificial intelligence AI. Bringing another industrial revolution. Previous ~~doctrine~~ revolutions expended humans mechanical power. This new revolution this second machine age. Is going to expand our cognitive. Abilities are mental power. Computers ~~are~~ not just going to replace manual labor but also ~~mentally~~ ~~for~~.

Test 4: Sprecher mit weißem Rauschen (laut)

Our world is changing in many ways. And one of the things. Which is ~~a hobby you in fact far~~ future artificial intelligence. A I. Bringing another ~~doctor were pollution. We do not remove.~~ Expanded human mechanical power. This new revolution. This second machine ~~a~~. Is going to expand our ~~call me to~~. Abilities ~~are~~ mental ~~problem~~. Computers are not ~~15 days~~ manual labor. But also ~~mentally~~.

Test 5: Sprecher mit instr. Musik (leise)

Our world is changing in many ways and one of the things which is ~~going to~~ have a huge impact on our future is artificial intelligence AI. Bringing ~~in~~ another ~~industry~~ revolution. Previous ~~doctrine~~ revolution expended humans mechanical power. This new revolution this second machine ~~a~~. Is going to expand our cognitive. ~~He's~~ our mental power. Computers are not just going to replace manual labor that also ~~mentally~~.

Test 6: Sprecher mit instr. Musik (laut)

Our world is changing in many ways and one of the things. Which is ~~going to~~ have a huge impact on our future is ~~official~~ ~~Intel~~. A I. ~~Other~~. .. ~~Did you~~. This. She. Are. Not just. ~~So~~.

Test 7: Sprecher mit Musik (leise)

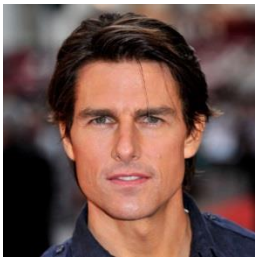
Our world is changing in many ways and one of the things which is ~~going to~~ have a huge impact on our future is artificial intelligence AI. Bringing another industrial revolution. ~~We've done 3 revolution~~ expended humans mechanical power. This new revolution. Second machine ~~a~~. Is going to expand our ~~call me to~~. Our mental power. Computers are not just going to replace manual labor but also ~~mentally~~.

Test 8: Sprecher mit Musik (laut)

~~Are we. The~~ way. One. ~~Just~~ gonna. ~~In fact. These~~ artificial. Another. ~~Daschle~~ revolution. ~~Expanded~~. ~~You~~.

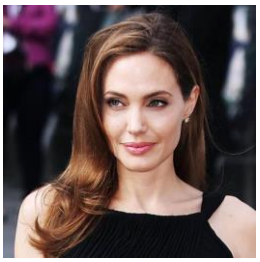
8. Bilder für IBM Watson Visual Recognition

1.) Tom Cruise



2.) Leonardo DiCaprio

3.) Angelina Jolie



4.) Heidi Klum

5.) Brad Pitt

6.) Veronica Ferres

7.) Heike Makatsch

8.) Til Schweiger

9.) Matthias Schweighöfer

10.) Nora Tschirner

9. Microsoft Azure: AI und Cognitive Services

Bildanalyse:

- **Maschinelles Sehen:** Infos aus Bilder filtern
- **Gesichtserkennung**: Gesichter auf Fotos erkennen, analysieren, organisieren und markieren
- **Content Moderator:** aut. Bild-, Text-, Videomoderation
- **Emotions**: Benutzererlebnis personalisieren mit Funktionen zur Erkennung der Emotion
- **Videoindexierung**: Infos aus Videos erschließen
- Benutzerdefinierter Dienst für maschinelles Lernen

Spracheingabe:

- **Sprachübersetzung**: Sprachübersetzungen in Echtzeit mit einem einfachen REST-API-Aufruf
- **Spracherkennung**: Identifizieren und authentifizieren Sie Sprecher anhand Ihrer Stimme
- **BING- Spracheingabe**: Umwandlung von Sprache in Text um und umgekehrt
- Benutzerdefinierter Dienst für maschinelles Lernen

Sprache:

- **Language Understanding Intelligent Service:** Befehle der Benutzer verstehen lernen
- **Textanalyse**: Stimmungen und Themen auswerten, um Benutzerwünsche zu verstehen
- **BING-Rechtschreibprüfung**: Ermittlung und Korrektur von Rechtschreibfehlern in der App
- **Textübersetzung**: maschinelle Übersetzung mit einem einfachen REST-API-Aufruf
- **Web Language Model**: Vorhersagesprachmodelle
- **API für linguistische Prozesse**: Vereinfachung von Sprachkonzepten und Textanalysen

Einblicke und Wissen

- **Empfehlung**: Prädiktion der Kundeninteressen mit Empfehlung dieser
- **Academic Knowledge**: Zugriff auf akademische Inhalte von Microsoft Academic Graph
- **Knowledge Exploration Service**: Interaktive Suche in strukturierten Daten über nat. Sprache
- **Q&A-Maker**: Umwandlung von Informationen in einfach zu findende Antworten

10. Ergebnisse Azure BING-Spracheingabe

Test 1: Sprecher

Our world is changing in many ways ~~than~~ one of the things which is gonna have a huge impact on our future is artificial intelligence, AI, bringing another industrial revolution. Previous industrial revolutions expended humans mechanical power. This new revolution this second machine age is going to expand our cognitive abilities ~~are~~ mental power. Computers are not just going to replace manual labor, but also mental labor.

Test 2: Sprecher mit Straßenatmo

Our world is changing in many ways ~~than~~ one. Artificial intelligence. Bringing another industrial revolution. Expended humans mechanical power. This is ~~Lucia~~ this second machine age is going to expand our cognitive abilities ~~are~~ mental power. ~~I'm~~ not just going to replace manual labor, but also mentally.

Test 3: Sprecher mit weißem Rauschen (leise)

Our world is changing in many ways and one of the things which is gonna have a huge impact on our future is artificial intelligence, AI, bringing another industrial revolution. Previous ~~doctor~~, ~~Aleutians~~ expended humans mechanical power. This ~~you~~ ~~Lucien~~ this second machine age is going to expand our cognitive abilities ~~are~~ mental power. ~~Butterz~~ are not just going to replace manual labor, but also mental labor.

Test 4: Sprecher mit weißem Rauschen (laut)

I will is changing in many ways and one of the things ~~that you can have it used in fact that she sure is~~. Bringing another industrial revolution. Expended humans mechanical power. This second machine ~~it~~. ~~He's~~ going to expand our cognitive abilities ~~are~~ mental ~~powers~~. ~~I'm~~ not just going to ~~be late 90 early~~ but also ~~mentally~~.

Test 5: Sprecher mit instr. Musik (leise)

Our world is changing in many ways and one of the things which is gonna have a huge impact on our future is artificial intelligence, AI, bringing another industrial revolution. ~~Microphallus. When~~

~~did~~ humans mechanical power? ~~You~~ ~~Lucien~~ this second machine age is going to expand our cognitive abilities ~~are~~ mental power. ~~I'm~~ not just going to replace manual labor, but also mental ~~E~~.

Test 6: Sprecher mit instr. Musik (laut)

Our world is changing in many ways and one of the things ~~just got here. Second shift. So~~

Test 7: Sprecher mit Musik (leise)

Our world is changing in many ways and one of the things which is gonna have a huge impact on our future is artificial intelligence. Bringing another industrial revolution. ~~Flu shots~~ expended humans mechanical power. ~~You Lucien~~ this second machine. ~~He's~~ going to expand our ~~colleges~~ ~~are~~ mental power. ~~I'm~~ not just going to replace manual labor, but also ~~mentally~~.

Test 8: Sprecher mit Musik (laut)

~~I will change it~~. Bringing another. ~~Expected. so~~