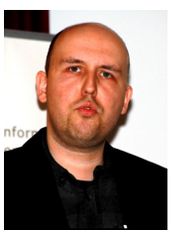
Integration von Audiomining in die Systemlandschaft von Radio Bremen

Oliver Hofrichter



Oliver Hofrichter
Radio Bremen
Diepenau 10
28195 Bremen
oliver.hofrichter@
radiobremen.de

Manuskript zum Vortrag anlässlich des Marianne-Englert-Preises auf der vfm-Frühjahrstagung am 10. April 2018.

EINLEITUNG / PROJEKTZIEL

Für die Erschließung von Essenzen spielen (semi-) automatische Verfahren eine zunehmend größere Rolle. Grund hierfür ist einerseits der technische Fortschritt, der Anwendungen ermöglicht, an die vor einigen Jahren noch nicht zu denken war. Andererseits sind die Rundfunkanstalten zunehmend von Kosteneinsparungen betroffen, wodurch in vielen Bereichen Effizienzsteigerungen erforderlich sind. Textbasierte Automatisierungsverfahren, wie Textmining sind in der ARD bereits seit mehreren Jahren im Einsatz, audiobasierte Verfahren, wie Audiomining hingegen noch nicht flächendeckend. Audiomining macht Sprachinhalte von unbekanntem Audiomaterial einer textbasierten Recherche zugänglich. Die Anwendung dieses Verfahrens wurde in den vergangenen Jahren beim Westdeutschen Rundfunk (WDR) erprobt und ist dort seit 2016 im produktiven Einsatz. Im Rahmen der Entwicklung des ARD-weiten crossmedialen Mediendatensystems (medas) wird die Audiomining-Lösung des Fraunhofer-Instituts für Intelligente Analyse- und Informationssysteme (IAIS) in die ARD-Hörfunkdatenbank (HFDB) integriert. Radio Bremen möchte Audiomining im Archiv-Kontext und später auch im Produktions-Kontext für Hörfunk und Fernsehen nutzen. Im Fokus des hier beschriebenen Projektes steht die Nutzung von Audiomining im Archiv-Kontext.

Die Abteilung Programmvermögen & Informationsservice (P.I.) archiviert ausgewählte Sendungen und Beiträge des Hörfunk- und Fernsehprogramms von Radio Bremen. Die Auswahl des zu archivierenden Contents erfolgt sowohl nach Absprache mit den Redaktionen, als auch auf Basis inhaltlicher Kriterien und kultureller Verpflichtungen. Beschränkte Personalressourcen haben zur Folge, dass nur ein Teil des Programmvermögens archiviert werden kann und es zu Unterschieden in der Tiefe der Erschließungen und damit auch in der Wiederauffindbarkeit der Inhalte kommt. Radio Bremen hat sich das Ziel gesetzt, den Anteil archivierter Programminhalte zu erhöhen, die Konsistenz innerhalb der Erschließungen zu steigern und die Auffindbarkeit von archiviertem Material zu verbessern. Hierfür soll Audiomining eingesetzt werden. Ein zentraler Bestandteil für die Nutzung von Audiomining im Archiv-Kontext ist die Integration dieser Technologie in die Archivdatenbank HFDB. Für diese Integration ist Radio Bremen der Pilotpartner innerhalb der ARD.

WAS IST AUDIOMINING?

Der englische Begriff „mining“ bedeutet Bergbau. Während beim Bergbau der Boden erkundet wird, werden beim Datamining Daten untersucht. Datamining im Allgemeinen bezeichnet die automatische Extraktion von Mustern aus Daten. Audiomining ist eine spezielle Form von Datamining. Der vom lateinischen „audire“ abgeleitete Namensbestandteil für hören oder zuhören deutet darauf hin, dass sich Audiomining auf Datenbestände konzentriert, die akustische Informationen enthalten. Bei Audiomining geht es also darum, Hörbares zu Tage zu fördern und Sprachinhalte von unbekanntem Audiomaterial einer textbasierten Recherche zugänglich zu machen (Beckers 2012). Der Fokus liegt hierbei auf der Verarbeitung von Wortbeiträgen. Audiomining-Funktionalität wird von verschiedenen Softwaresystemen bereitgestellt. Die von der ARD und auch im Rahmen dieses Projektes eingesetzte Software von Fraunhofer bietet folgende Funktionalitäten (Schmidt et al. 2016):

- Segmentierung: das Audiosignal wird automatisch in Segmente unterteilt, die jeweils das gesprochene Wort eines individuellen Sprechers enthalten
- Erkennung sprachlicher und nichtsprachlicher Anteile (Musik) des Audiosignals
- Erkennung des Geschlechts der sprechenden Personen

- Sprecherclustering / Sprechertracking: innerhalb der Struktureinheiten werden wiederkehrende Sprecher erkannt
- Automatische Spracherkennung (automatic speech recognition, ASR): Generierung eines Transkripts (MPEG-7) des gesprochenen Wortes
- Keyword-Extraktion: Extrahieren der wichtigen sinntragenden Begriffe aus dem Audio

Eine der zentralen Komponenten innerhalb eines Audiomining-Systems ist die automatische Spracherkennung. Sie realisiert die Umwandlung von gesprochener Sprache in geschriebenen Text und basiert auf statistischen Modellen, die trainiert werden. Diese kommen im Anschluss an die Signalverarbeitung des zu analysierenden Audioinhalts zum Einsatz. Für die Dekodierung der verarbeiteten Signale werden folgende Modelle eingesetzt: akustisches Modell, Lexikon und Sprachmodell. Das Sprachmodell modelliert die Wahrscheinlichkeit von Wortfolgen. Das Lexikon definiert eine Liste von Wörtern, die erkannt werden sollen, in Kombination mit ihrer Aussprache. Das akustische Modell repräsentiert den Klang der Phoneme der jeweiligen Sprache in einem maschinenlesbaren Format. Die Abfolge der einzelnen Analyseschritte zeigt Abbildung 1.

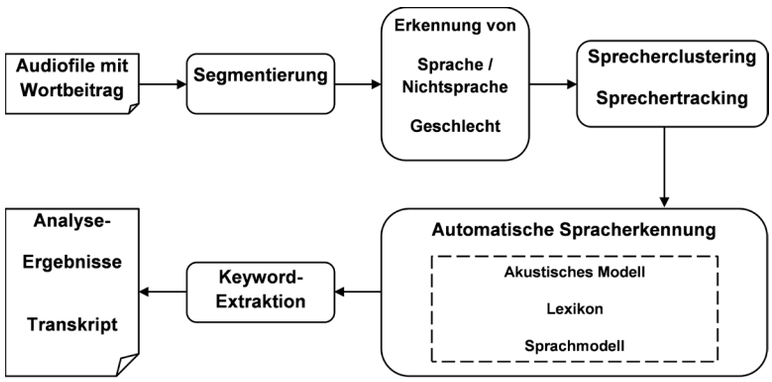


Abbildung 1

EVALUATION

Um eine technische Lösung wie Audiomining optimal in bestehende Geschäftsprozesse einbinden zu können, ist es wichtig, die Stärken dieser Technologie voll auszuschöpfen und sich ihrer Schwächen bewusst zu sein. Zur Identifikation von Stärken und Schwächen von Fraunhofer IAIS Audiomining wurde Radio Bremen für dieses Projekt im Rahmen eines Auftrags an Fraunhofer ein Demosystem zur Verfügung gestellt. Darüber hinaus gab es einen regen Wissens- und Erfahrungsaustausch mit Fraunhofer und WDR und seit Mitte 2017 mit weiteren Rundfunkanstalten innerhalb der zu diesem Zeitpunkt neu gegründeten ARD-Expertengruppe Audiomining.

Im Interesse einer längerfristigen Entwicklung und Verbesserung von Audiomining und zur Vermeidung einer ausschließlich subjektiven Evaluation des Systems, wurde Radio Bremen kurz nach Bereitstellung des Audiomining-Demosystems von Fraunhofer gebeten, Testdaten zu erstellen. Vorgabe für dieses sogenannte annotierte Sprachtestset war, Sendungen im Gesamtumfang von etwa einer Stunde zusammenzustellen. Da Radio Bremen Audiomining sowohl für den Hörfunk, als auch für das Fernsehen einsetzen möchte, wurde für beide Bereiche Content von jeweils einer Stunde Gesamtdauer zusammengestellt. Um keine willkürliche Auswahl vorzunehmen, wurde ausgewertet, welcher Content im Jahre 2016 von Radio Bremen archiviert wurde. Auf Basis dieser

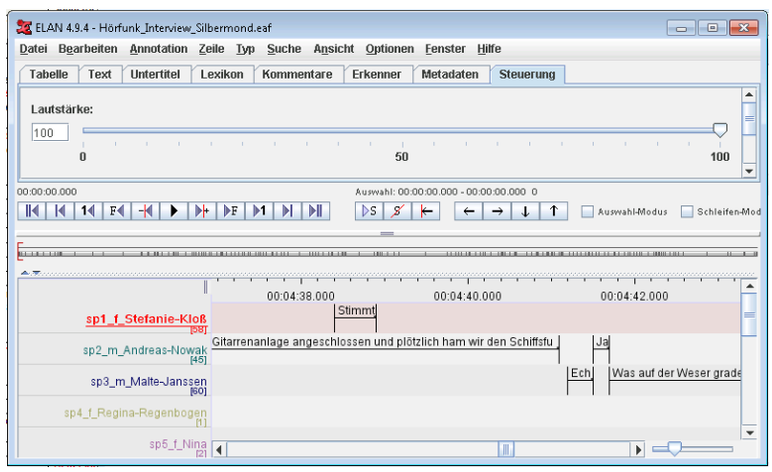


Abbildung 2

Informationen wurde entsprechend der jeweiligen zeitlichen Anteile im Vergleich zur archivierten Gesamtmenge, Content ausgewählt, der in diesem Sinne typisch für Radio Bremen ist. Weitere Vorgabe seitens Fraunhofer war, dass das annotierte Sprachtestset mit der Software ELAN (EUDICO Linguistic Annotator) zu erstellen ist. Die Abbildung 2 zeigt einen Screenshot aus dieser Software. Gezeigt wird dort ein Auszug eines annotierten Transkripts eines Hörfunkinterviews der Welle Bremen Vier mit der Band Silbermond.

Annotationen in ELAN sollten nach dem folgenden Schema erfolgen:

- Jeder Sprecher wird auf einer separaten Ebene annotiert.
- Benennungsschema: „sp+laufende Nummer+
- Geschlecht+Name“,
- also z.B. „sp1_f_Angela-Merkel“ (sp = engl. Speaker)
- In jedem Annotationslayer werden die gesprochenen Wörter zeitgenau annotiert.
- Neben reinem Hochdeutsch sind auch umgangssprachliche Ausdrücke, wie z.B. „ham“ (haben), „ne“ (eine), „se“ (sie), erlaubt.
- Auch Verzögerungslaute (Hesitationen), wie „äh“, „ähm“ oder „mhh“ sollen transkribiert werden.
- Zahlen dürfen als Ziffern oder textuell ausgeschrieben werden.
- Zusammengefasst sind die Ziele dieses Arbeitspaketes:

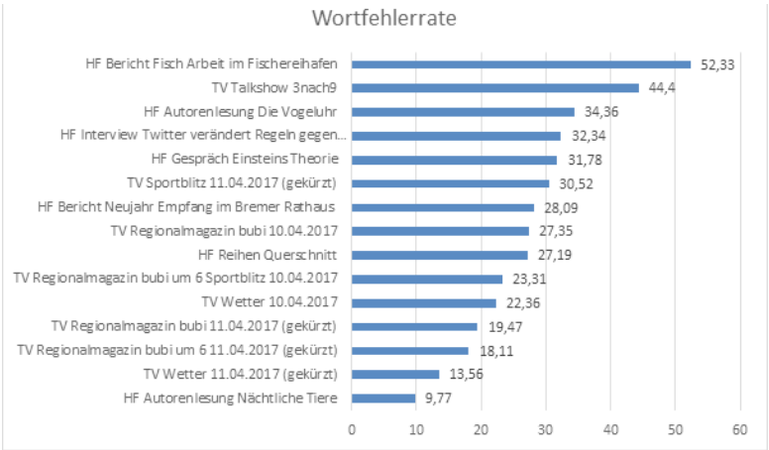
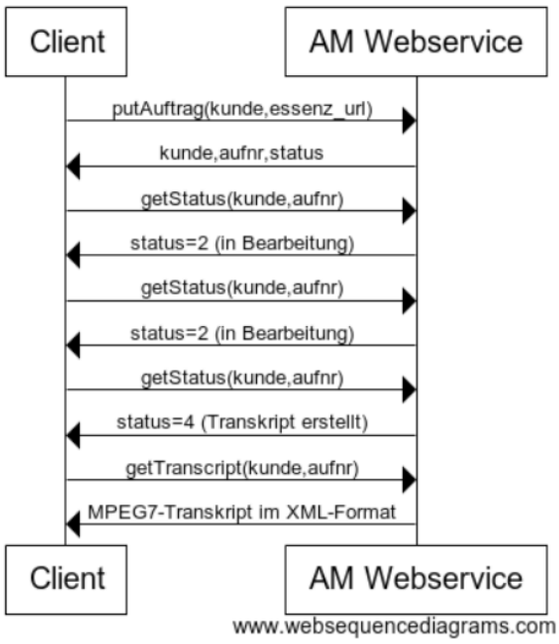


Abbildung 3
Abbildung 4



- Bestimmung belastbarer, objektiver Zahlen im Sinne einer Wortfehlerrate
- Sprecherclustering / Sprechertracking testen
- Sprechererkennung / -identifikation für relevante Sprecher testen
- Sprechersegmentierung (Evaluation durch Zeitgrenzen der Annotation)

Auf Basis der von Radio Bremen gelieferten Daten berechnete Fraunhofer Fehlermaße, wie die der Abbildung 3 zu entnehmende Wortfehlerrate. Hörfunkinhalte sind dabei durch das Präfix „HF“ gekennzeichnet, Inhalte aus dem Fernsehen mit „TV“: Die höchste Wortfehlerrate weisen ein Hörfunkbericht aus dem Bremerhavener Fischereihafen und der Audiomitschnitt der Radio-Bremen-TV-Talkshow 3nach9 auf. Der Bericht aus dem Fischereihafen enthält vor allem O-Töne in norddeutschem Dialekt und viele Hintergrundgeräusche. Die Audiospur der Talkshow bereitet Audiomining vor allem deshalb Probleme, da häufig viele Menschen gleichzeitig und durcheinander sprechen und Hintergrundgeräusche wie Applaus vorkommen. Die geringste Wortfehlerrate zeigt sich bei einer Autorenlesung. Ein Querschnitt verschiedener Hörfunkreihen Radio Bremens bewegt sich mit einer durchschnittlichen Fehlerrate von 27 % im Mittelfeld. Die Audiomitschnitte aus dem Fernsehbereich bewegen sich von der Fehlerrate her im unteren Bereich.

WORTDOKUMENTATION BEI RADIO BREMEN

Radio Bremen betreibt mit Bremen Eins, Bremen Zwei, Bremen Vier sowie Bremen NEXT vier Hörfunkwellen und liefert Programmteile für die Welle COSMO zu. Täglich werden rund 70 Stunden Hörfunk-Programm gesendet. Der Wortanteil darin liegt bei ca. 25 Stunden täglich. Derzeit werden aus dem gesendeten Programm täglich im Durchschnitt etwa 3 ½ Stunden an Wortbeiträgen archiviert. Dies entspricht ca. 15 % des Wortanteils. In der ersten Phase der Audiomining-Nutzung (seit April 2018) wird Audiomining im Rahmen der regulären täglichen Archivierung von Wortbeiträgen eingesetzt. Zu einem späteren Zeitpunkt der Audiomining-Nutzung soll das Angebot archivierter Inhalte vergrößert und Audiomining auch (automatisch) aus dem Hörfunk-Produktionssystem d‘accord angestoßen werden. Die Archivierung von Wort-Content erfolgt größtenteils auf Basis eines Lektorats des täglichen Sendeplans. Dem Lektorat liegen inhaltliche Kriterien und kulturelle Verpflichtungen zu Grunde: unter Anwendung der Archivierungskriterien werden in d’accord in den Ist-Sendeplänen der einzelnen Hörfunkwellen archivierungswürdige Inhalte identifiziert. Nach der Auswahl der Inhalte wird ein Grundgerüst an Metadaten für den Transfer in die Archivdatenbank HFDB vorbereitet. Der eigentliche Transfer erfolgt mittels hauseigener Plugins. Der Prozess wird durch die inhaltliche Erschließung in der HFDB abgeschlossen.

INTEGRATION

Die Audiomining-Funktionalität wird als SOAP-Webservice bereitgestellt. Die eigentliche Analyse wird durch einen Analyseserver bewerkstelligt, dem ein Auftragsvergabesystem (AVS) vorgeschaltet ist. Konkret stehen drei Webservice-Methoden zur Verfügung:

- Beauftragung
- Status-Abfrage
- Abholung der Ergebnisse

Die Beauftragung von Audiomining inklusive Abholung der Analyseergebnisse erfordert den sequenziellen Aufruf dieser drei Methoden. Zunächst wird im ersten Schritt eine Analyse beauftragt. Hierzu wird die URL einer Essenz aus dem Essenzsystem der jeweiligen Rundfunkanstalt übermittelt. Das

Auftragsvergabesystem vergibt für den Auftrag eine ID. Sobald der Auftrag im Auftragsvergabesystem an der Reihe ist, wird ein Audiominingservice mit der Bearbeitung beauftragt und greift über die URL auf die Essenz zu. Das anfragende System kann nun mittels Status-Abfrage Informationen über den Fortschritt der Analyse abrufen. Nach erfolgreicher Erstellung des Transkriptes kann dieses über die Auftrags-ID abgerufen werden. Das UML-Sequenzdiagramm in Abbildung 4 veranschaulicht exemplarisch die Beauftragung von Audiomining inklusive Abholung der Analyseergebnisse. Ab der Version 9 nutzt die HFDB diese Schnittstelle zur Realisierung ihrer neu eingeführten Audiomining-Funktionalität. Die Funktionalität umfasst im Einzelnen:

- Steuerung des Audiominingprozesses
- Darstellung der Audiomining-Analyseergebnisse
- Möglichkeiten der Recherche nach Audiomining-Analyseergebnissen
- Nutzung der Audiomining-Analyseergebnisse im Player

Audiomining lässt sich ausschließlich aus dem HFDB-Erfassungs-Client (Rich-Client) heraus auslösen, redaktionelle Nutzer des Web-Clients können den Prozess nicht initiieren. Die Funktionalität zum Starten des Audiominingprozesses wird über ein HFDB-Plugin bereitgestellt. Audiomining kann über eine Schaltfläche in verschiedenen Ansichten innerhalb der HFDB (Kurzinfo, Vollinfo und Merkliste) ausgelöst werden. Daraufhin wird eine Audiomining-Beauftragung durchgeführt. Ihr Status lässt sich in der Workflowsteuerung der HFDB verfolgen. Nach erfolgreicher Verarbeitung lassen sich die Analyseergebnisse in verschiedenen Ansichten innerhalb der Vollinfo von Web- und Rich-Client sowie der Erfassung betrachten. Sowohl Web-, als auch Rich-Client bieten Recherchemöglichkeiten für Transkript und Keywords und ermöglichen ebenfalls den Zugriff auf die Analyseergebnisse mittels neuem Audioplayer.

Als Pilotpartner für die Integration von Audiomining in die HFDB war Radio Bremen für das Qualitätsmanagement aller HFDB-Funktionalitäten rund um das Thema Audiomining verantwortlich.

Die technische Integration von Audiomining bei Radio Bremen begann mit der Nachrüstung einer Vorhörmöglichkeit für die eigene HFDB-Testinstanz. Die Vorhör-Funktionalität hatte zuvor ausschließlich in der Produktivumgebung zur Verfügung gestanden. Auch das Vorhörformat musste an die Anforderungen der neuen HFDB-Version angepasst werden. Außerdem wurden die Plugins, die bei Radio Bremen die HFDB mit dem Produktionssystem d'accord vernetzen, um Audiomining-Funktionalität erweitert, sodass während des Archivierungsvorgangs beim Metadaten-Transfer von d'accord nach HFDB

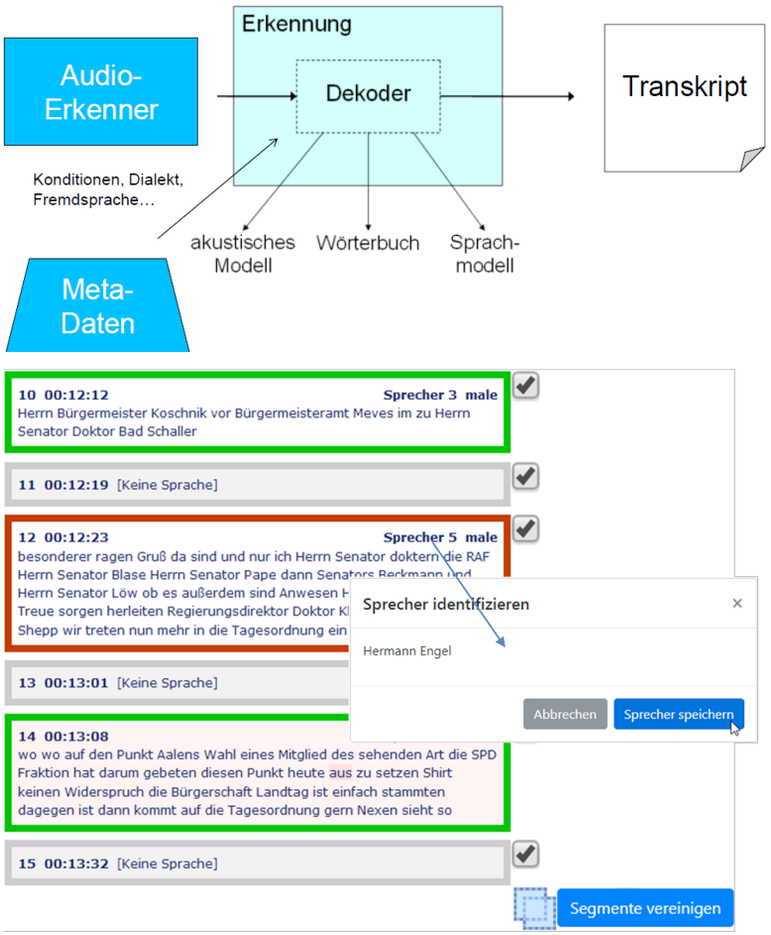


Abbildung 5
Abbildung 6

automatisch ein Audiomining-Auftrag für die jeweilige Essenz getriggert wird.

MÖGLICHKEITEN, GRENZEN UND OPTIMIERUNGSVORSCHLÄGE

Automatische Spracherkennung verwendet eine statistische Repräsentation der gesprochenen Sprache in Form von Modellen. Während das Sprachmodell Informationen darüber enthält, welche Wortfolgen im Anwendungskontext zu erwarten sind, ist sein akustisches Pendant ein Modell der Geräuschmuster, aus denen einzelne Wörter bestehen. Die Modelle werden trainiert und können nur mit in ihnen enthaltenen Begriffen umgehen. In den Medien tauchen häufig neue Begrifflichkeiten und Namen auf. Die Modelle müssen also auf dem neuesten Stand gehalten werden. Dies wird in Zukunft vollautomatisch auf Basis von aus dem Internet gecrawlten Texten erfolgen. Hierfür ebenfalls denkbar ist die Nutzung von Manuskripten, die direkt aus den Rundfunkanstalten stammen und exakt die im Programm verwendete Sprache nutzen. Bei der Auswahl des Textkorpus ist sicherzustellen, dass die regionale Spezifität der Sprachmodelle gewährleistet ist. Es muss also pro Rundfunkanstalt möglich sein, Quellen für eigene Sprachmodelle zu pflegen. Hierfür müssen auch personelle Ressourcen eingeplant werden.

Für das Erlernen der Aussprache neuer Begriffe muss das akustische Modell aktualisiert werden. Hierfür ist eine Anbindung der ARD-Aussprachedatenbank geplant.

Aktuell erkennt Fraunhofer IAIS Audiomining ausschließlich deutsche Sprache. Fremdsprachige Inhalte werden nicht als solche identifiziert. Sobald das System auf einen fremdsprachigen Inhalt stößt, wird diejenige deutsche Wortfolge transkribiert, die aus der Kombination der drei Modelle als die wahrscheinlichste hervorgeht. Dadurch gelangt an solchen Stellen „Abfall" in die Transkripte. Diese Problematik kann zweistufig bearbeitet werden. Ein erster Fortschritt bestünde bereits darin, fremdsprachige Inhalte als solche zu identifizieren und diese Segmente nicht zu transkribieren. In einem zweiten Schritt könnten die Modelle so angepasst werden, dass zunächst neben deutschen speziell auch englische, später dann zusätzlich anderssprachige Inhalte transkribiert werden können. Die Berücksichtigung von Fremdsprachen und Dialekten könnte durch einen im Erkennungsprozess vorgelagerten Audio-Erkenner (siehe Abbildung 5) realisiert werden. Dieser würde die Audiokonditionen analysieren und nachgelagert die jeweils passenden Modelle für die Weiterverarbeitung aktivieren. Auch Spontansprache könnte auf diese Weise berücksichtigt werden.

Eine weitere Optimierung bestünde in der Lernfähigkeit des Systems: In Abbildung 6 wird skizziert, wie Nutzer während dokumentarischer Tätigkeiten in der Transkript-Ansicht Optimierungen hinsichtlich Segmentierung oder Sprecheridentifikation vornehmen könnten: falsch segmentierte Bereiche könnten zunächst markiert und anschließend per Mausklick vereinigt werden, nicht oder falsch erkannte Sprecher nachträglich manuell identifiziert werden. Letzteres könnte eine Übertragung des i-Vector (aus rund 400 Merkmalen bestehender Hashwert, der einen Sprecher charakterisiert) in den Personen-Datensatz einer zentralen Datenbank, wie der Normdatenbank (NDB), auslösen.

Für die Sprecheridentifikation ist es wichtig, dass es einerseits einen gemeinsamen Pool an Sprecherinformationen gibt, andererseits aber jede Rundfunkanstalt auch individuell ihre eigenen Sprecher pflegen kann, um beispielsweise die Identifikation regionaler Politiker oder der eigenen Reporter gewährleisten zu können. Um einen Eindruck von der Leistungsfähigkeit der Sprecheridentifikation zu bekommen, hat Radio Bremen Fraunhofer zwei Sendungen des Regionalmagazins „buten un binnen" zur Verfügung gestellt, in denen Reporter sprechen, die auch im manuell annotierten Sprachtestset vorkommen. Der zeitliche Umfang der Trainingsdaten für diese Sprecher bewegte sich im zeitlichen Umfang von einer halben bis knapp 8 Minuten. In dieser Stichprobe konnte kein direkter Zusammenhang zwischen Umfang des Trainingsmaterials und Erkennungsrate des jeweiligen Sprechers festgestellt werden. Auch Sprecher, zu denen wenig Trainingsmaterial vorlag, wurden überdurchschnittlich gut erkannt. Da eine Maschine im Kontext der Sprecheridentifikation zum Teil durchaus auch Merkmale, wie Hintergrundgeräusche lernt, war bei einem Sprecher X, der im Trainingsmaterial häufig bei starken Hintergrundgeräuschen sprach, eine schlechte Erkennungsquote zu beobachten. Die Schlussfolgerung der Maschine: sobald ein Segment viele Hintergrundgeräusche enthält, spricht dort Sprecher X. Ob die Sprecheridentifikation in ihrem jetzigen Entwicklungsstadium bereits reif für den Produktivbetrieb ist, werden (weitere) Proofs of Concept zeigen.

FAZIT UND AUSBLICK

Mit der erfolgreichen Integration von Audiomining in die Systemlandschaft von Radio Bremen wurde das Projektziel erreicht. Radio Bremen setzt Audiomining seit April 2018 produktiv ein. Neben der konkreten technischen Integration wurde ein Konzept für die Integration von Audiomining in das bei Radio Bremen eingesetzte Hörfunk-Produktionssystem entwickelt. Analysen, wie eine Ist-Analyse der Wortdokumentation bei Radio Bremen, eine Betrachtung des möglichen Zusammenspiels von Audiomining und Textmining, eine Untersuchung alternativer Audiomining-Lösungen sowie der Auswirkungen von Audiomining auf bestehende Geschäftsprozesse, rundeten das Projekt ab.

Die Einführung eines automatischen Verfahrens führt zu Veränderungen innerhalb von Geschäftsprozessen. Audiomining ist da keine Ausnahme. Es wird zu Veränderungen in der Wort-, später auch in der Videodokumentation führen und sich auch auf den Redaktionsalltag auswirken: Die Bedeutung von (Audio-)Mining für Nicht-Archivsysteme wird, gerade in Zeiten stetig wachsender Mengen an unstrukturierten Daten bei gleichzeitiger Ressourcenknappheit, zunehmen. Automatische Verfahren zur Auswertung von Inhalten werden Dokumentation und Redaktionen bei der Erzeugung hochwertiger Informationsprodukte unterstützen. Eine Live-Transkription während einer Pressekonferenz, an die sich direkt ein Schnitt des Audio-Rohmaterials auf Basis des generierten Transkriptes anschließt, ist da nur ein mögliches Zukunftsszenario.

Literatur

[Beckers 2012] Beckers, Thomas (2012): Audiomining – Noch „Hexenküche" oder bereits Werkstatt?, in: info7, Jg. 27, Nr. 3, S. 13-17.

[Schmidt et al. 2016]
Schmidt, Christoph; Stadtschnitzer, Michael; Köhler, Joachim (2016): The Fraunhofer IAIS Audio Mining System: Current State and Future Directions, in: 'Proceedings of the 12. ITG Symposium on Speech Communication, Paderborn, Germany, October 5-7, 2016', IEEE.