Arjan van Hessen · Franciska de Jong · Stef Scagliola

Der Einsatz von Sprachtechnologie in Oral-History-Sammlungen

Aus dem Englischen übersetzt von Heiko Pollmeier

Oral-History-Aufnahmen – audiovisuelle ebenso wie reine Höraufnahmen – haben die Einschränkung, ohne schriftliche Darstellung nicht durchsuchbar zu sein. Man kann ihren Inhalt nur dann analysieren, wenn man die komplette Aufnahme anschaut bzw. anhört und Notizen macht. Diese Vorgehensweise ist zeitaufwendig und schränkt die bearbeitbare Interviewmenge stark ein. Mit der wachsenden Zahl verfügbarer Oral-History-Sammlungen, von denen die meisten keine durchsuchbaren Transkriptionen enthalten, ist die Datenmenge nicht mehr länger durchsuchbar noch analysierbar. Indem man den Aufnahmen umfassende Metadaten wie beispielsweise Zusammenfassungen (kurze, die auf die Gesamtaufnahme Bezug nehmen, oder segmentbezogene für 10-Minuten-Abschnitte der Aufnahme) sowie Schlagwörter hinzufügt, oder indem man den Inhalt mittels Kapiteln strukturiert, hilft man den Nutzern, durch die Aufnahme zu navigieren. Dennoch kostet es weiterhin viel Zeit, spezifische Themen, Wörter oder Aussagen herauszufiltern.

Das hier angerissene Problem ist weder für Oral-History-Sammlungen noch für audiovisuelle Daten überhaupt spezifisch. Große Archivsammlungen sind immer arbeitsintensiv und bedürfen zwangsläufig manueller Arbeit, um sie zu erschließen. Und doch ist die Entwicklung von Textsuchmaschinen ein großer Fortschritt. Das Suchen und Herausfiltern von Informationen ist heutzutage so leistungsstark, dass Millionen Dokumente unmittelbar durchsucht werden können. Eine wesentliche Hilfe bei der Erschließung von Oral-History-Sammlungen in großem Rahmen liegt also darin, gesprochene Sprache in eine textuelle, verschriftlichte Darstellung umzuwandeln, die dann mit eigens für das Textmaterial entworfenen Systemen bearbeitet und durchsucht werden kann.

Neue Möglichkeiten auf dem Gebiet der Human Language Technology (HLT)[1] und die wachsende Menge an digitalen Oral-History-Aufnahmen stimulieren den Einsatz von HLT-Tools zum Aufbau interaktiver Plattformen mit Direktzugang zu Oral-History-Daten. Der vorliegende Beitrag stellt die Sprachtechnologie in den Mittelpunkt, und zwar insbesondere die Spracherkennung und ihr Potenzial, Aufnahmen bequemer und schneller zu bearbeiten und auf der Wortebene zugänglich zu machen. Diverse Fälle mit unterschiedlichen Spracherkennungsmethoden werden vorgestellt und ihre potenziellen Vorzüge illustriert.

1 „Human Language Technology" (HLT) oder „Natural Language Processing" (NLP) ist ein Forschungsfeld auf dem Gebiet der Computerwissenschaften, Künstlichen Intelligenz und der Linguistik, das sich mit der Interaktion zwischen Computern und den (natürlichen) menschlichen Sprachen beschäftigt.

1. Automatische Spracherkennung

Idealerweise vermag die Automatische Spracherkennung (engl.: Automatic Speech Recognition, ASR), gesprochene Inhalte in Text umzuwandeln. Leider sind die ASR-Maschinen zurzeit noch nicht ausreichend effizient, um fehlerfreie Transkriptionen des in alltäglichen, nicht fachspezifischen Konversationen Gesagten hervorzubringen; für Niederländisch beispielsweise werden die Ergebnisse keine 60 Prozent[2] an korrekter Erkennung erreichen. Die Gründe für die geringe Leistung sind:

1. Bei vielen Oral-History-Sammlungen sind die Aufnahmebedingungen alles andere als ideal (während der Interviews muss man oft improvisieren).
2. Menschen sprechen nicht flüssig sowie grammatikalisch falsch (was fast immer der Fall ist).
3. Menschen sprechen Dialekt oder eine andere Sprache als ihre Muttersprache.

1.1 Arbeitsabläufe bei der Automatischen Spracherkennung
Die Arbeitsschritte bei der Automatischen Spracherkennung sind im Einzelnen:

1. Im Falle eines Videos wird die Sprache einer audiovisuellen Datei herausgefiltert, und die Tonspur wird in ein ASR-geeignetes Format, etwa eine wav-Datei, umgewandelt.
2. Das Sprachmodell für die jeweilige Sprache wird ausgewählt.
3. Das Sprachmodell wird um themenspezifische Wörterlisten und Texte ergänzt.

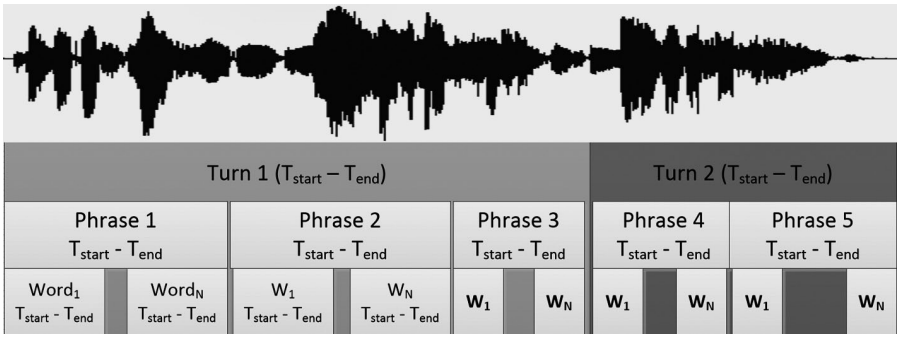
Das Ergebnis ist eine Liste mit Wörtern mit der Angabe von Start- und Endzeit. Unterstützt die ASR-Maschine Sprecherwechselerkennung (kann also bestätigen, dass eine andere Person spricht) und Sprecheridentifizierung (bestätigen, wer diese Person ist), kann beides den Erkennungsergebnissen zugefügt werden (vgl. Abbildung auf der nächsten Seite).

Eine automatisch erstellte Transkription in einer zur Untertitelung und zur Textanalyse geeigneten Qualität bleibt noch lange Jahre ein Traum. Jedoch können ASR-Ergebnisse erfolgreich zum Durchsuchen audiovisueller Archive genutzt werden, sogar bei einer Transkription, in der nur 60 Prozent der Wörter korrekt übertragen sind.[3]

1.2 Verbesserung der Automatischen Spracherkennung
An der Verbesserung von ASR arbeiten Fachleute von Universitäten und aus der Wirtschaft. Dennoch lässt sich auch ohne die Programmierung neuer Software die

2 Vgl. David van Leeuwen/Judith Kessens/Eric Sanders/Henk van den Heuvel, Results of the N-Best 2008 Dutch speech recognition evaluation, in: Proceedings of Interspeech (2009), S. 2531–2534.

3 Michael Levit/Shuangyu Chang/Bruce Buntschuh/Nick Kibre. "End-to-end speech recognition accuracy metric for voice-search tasks", in: Proceedings of International Conference on Acoustics, Speech & Signal Processing (ICASSP) (2012), S. 5141–5144.



Darstellung einer Sprachwelle: Die ASR-Maschine hat zwei Gesprächsschritte/Redebeiträge (*turn* = T) mit drei Äußerungseinheiten (*phrase*) in der ersten und zwei Äußerungseinheiten im zweiten Gesprächsschritt festgestellt. Jede Äußerungseinheit hat n Wörter (W), alle mit einem Beginn und einem Ende.

Leistung der Automatischen Spracherkennung verbessern. Ob ein solcher Aufwand lohnt, ist von Fall zu Fall zu entscheiden.

Automatische Spracherkennung basiert auf zwei Annahmen: a) Die Phoneme (Sprachlaute) eines individuellen Sprechers unterscheiden sich nicht wesentlich von den durchschnittlichen Phonemen, wie sie für jene spezifische Sprache gesammelt wurden, b) Die meisten vom Sprecher geäußerten Wörter und Wortkombinationen entsprechen dem statistischen Sprachmodell, das auf der Basis riesiger Textmengen erstellt worden ist. Spezifischere Sprachmodelle sind effizienter: Die ASR wird ein gesprochenes Dokument über die Farbgebung bei Rembrandt besser erkennen, wenn das Sprachmodell auf Schriftdokumenten über den Einsatz von Farben in der holländischen Malerei des 17. Jahrhunderts beruht – und nicht auf Dokumenten der Jahrestreffen des Internationalen Währungsfonds.

Bessere akustische Modelle, bessere Aussprachеprognosen und bessere Sprachmodelle können ASR-Maschinen verbessern helfen.

Akustische Anpassung

Jede Stimme ist anders; dennoch nutzen ASR-Maschinen ein akustisches Modell zur Erkennung der gesprochenen Sprache, das auf durchschnittliches Werten beruht. Am Anfang steht stets ein Modell pro Sprache. Unterscheidet sich jedoch die Aussprache in den zahlreichen geografischen Regionen, in denen jene Sprache gesprochen wird, stark, ist es sinnvoll, je eigene akustische Modelle zu entwickeln. Als die ersten ASR-Maschinen kommerziell genutzt wurden, schien die Aussprache von amerikanischem, britischem, indischem, südafrikanischem und australischem Englisch ausreichend unterschiedlich zu sein, um die Entwicklung verschiedener akustischer Modelle zu rechtfertigen. Für das Deutsche wären ein deutsch-deutsches, ein deutsch-schweizerisches sowie ein deutsch-österreichisches Modell denkbar. Eine weitere Verfeinerung könnte in einem geschlechtsspezifischen Modell bestehen: Frauen und Männer haben unterschiedliche Stimmwege und unterschiedliche Stimmbänder, was zu deutlich voneinander unterscheidbaren Stimmen

führt. Solche regionalen und genderspezifischen akustischen Modelle sind inzwischen Standard bei den meisten ASR-Maschinen. Das ultimative Akustikmodell wäre dasjenige eines individuellen Sprechers; allerdings ist es nicht realistisch, dies für alle Sprecher in einem Interviewprojekt umzusetzen. Anstatt ein komplett individualisiertes akustisches Modell zu entwickeln, kann man das existierende Modell anpassen, und zwar auf der Grundlage einer fünf- bis zehnminütigen Aufnahme einer Sprecherstimme.

Der gesprochene Text muss sorgsam transkribiert werden, um den Computer an die unterschiedliche Aussprache der Vokale und Konsonanten durch die jeweilige Person zu gewöhnen. Solche leicht anders gesprochenen Phoneme bilden die Grundlage für die Entwicklung eines individualisierten akustischen Modells. Der dafür nötige Zeitaufwand lohnt sich schon dann, wenn die Interviews länger als 30 Minuten sind.

G2P (Graphem->Phonem)

Sogenannte G2P-Konvertierer – G2P steht für „Grapheme-to-Phoneme", also die Umwandlung von Graphemen[4] in Phoneme, grob vereinfacht: die Umwandlung von Buchstabenfolgen in Lautfolgen – sind kleine Softwareprogramme, die Wörter phonetisch transkribieren. Ein moderner G2P-Konvertierer benutzt ein Wörterbuch und – für in diesem Wörterbuch nicht verzeichnete Wörter – einen Satz sprachabhängiger Regeln. Wenn beispielsweise Niederländer von ihrer Zeit in Deutschland während des Zweiten Weltkriegs erzählen, dann benutzen sie kriegsrelevante deutsche Wörter wie „Sturmbannführer". Ein G2P-Konvertierer für Niederländisch wird dieses Wort nicht in seinem Wörterbuch haben und zudem niederländische Transkriptionsregeln befolgen, die zu einer unsinnigen Transkription führen. Deswegen ist eine manuelle Transkription nötig. Das Hinzufügen besonderer, themenrelevanter Wortlisten (inkl. ihrer phonetischen Transkription) kann zu erheblich besseren phonetischen Transkriptionen und so zu besseren Erkennungsergebnissen führen.

Sprachmodellanpassung

Die Spracherkennung nutzt das akustische Signal, um einen aneinandergereihten Phonemstrom zu produzieren. Es ist jedoch praktisch unmöglich, die Wörter allein auf der Grundlage erkannter Phoneme zu produzieren. Man stelle sich vor, die eingegebene Phrase wäre „recognize speech" („erkenne Sprache"). Das phonetische Äquivalent wäre dann:

Wörter	Phonetische Transkription (SAMPA-Format)
recognize speech [Sprache erkennen]	r E k @ n Ai z p i: t s
wreck a nice beach [ruiniere einen schönen Strand]	

4 Ein Graphem ist die kleinste bedeutungsunterscheidende grafische Einheit in einem Schriftsystem, die ein Phonem repräsentiert.

Die Phonemketten werden mithilfe eines Sprachmodells in Ketten der wahrscheinlichsten Wörter umgewandelt: Ein statistisches Modell prognostiziert die Möglichkeit, dass das Wort C gesagt wird, und die Wahrscheinlichkeit, dass dieses Wort C auf die Wörter A und B folgt. Um diese Wahrscheinlichkeit zu berechnen, bedarf es einer riesigen Textmenge.

Gemäß der allgemeinen ASR-Methode wird so viel digital verfügbarer Text wie möglich benutzt, um die diversen Wahrscheinlichkeiten zu kalkulieren. Die Annahme lautet jedoch, dass diese Texte die Sprache wie in den Aufnahmen gesprochen darstellen. Das mag bei den 20-Uhr-Nachrichten zutreffen, jedoch nicht in den meisten Oral-History-Aufnahmen, wo Menschen über Ereignisse in der Vergangenheit tendenziell mit anderen, selten benutzten Wörtern sprechen. Um also die Wahrscheinlichkeit dieser gesprochenen Wörter korrekt zu berechnen, bedarf es eines Textes über jene Ereignisse. Ein Beispiel ist das deutsche Wort „Ostarbeiter". In den meisten modernen Texten kommt dieses Wort nicht vor, in Erzählungen ehemaliger Zwangsarbeiter allerdings oft. Der Rückgriff auf Texte über Zwangsarbeit erhöht die Möglichkeit, dass die ASR-Maschine das Wort „Ostarbeiter" erkennt.

1.3 Alignment

Eine spezielle Version der Spracherkennung ist das Alignment, also die Synchronisierung von Audio und Text bzw. die zeitliche Kopplung von Mediendatei und Transkript. Wie erwähnt, hängt die Präzision von ASR stark vom Sprachmodell ab: Je besser die Prognose, umso besser das Resultat. Idealerweise weiß man genau, WAS gesagt werden wird, und die ASR-Maschine muss nur noch das WANN ($= S_{Beginn}$ und S_{Ende}) erkennen. Dies ist der Fall bei einer (manuell erstellten) Transkription. Die Spracherkennung erhält die Tonspur sowie den gesprochen Text und muss herausfinden, wann welches Wort gesprochen wurde. Im Allgemeinen ist eine solche Angleichung eine sehr einfache, schnelle und daher verlässliche Aufgabe für ASR-Maschinen.

Probleme

Die drei häufigsten Störquellen für die Koppelung von Audiodateien und Texten sind:

1. Hintergrundlärm: Dieser muss während der Aufnahmezeit verhindert werden. Steht Geräuschunterdrückung nicht zur Wahl, ist die Koppelung schwierig. In diesem Fall kann man Markierungen setzen, zum Beispiel alle fünf Minuten. Das Angleichungsprogramm „weiß" dann, dass die Sprache zwischen Start- und End-Markierungen zu finden sein muss. So wurde bei der Angleichung der „Radio Oranje"-Sammlung mit 37 zwischen 1940 und 1945 übertragenen Reden der niederländischen Königin Wilhelmina vorgegangen (siehe unten).
2. „Seltsame" Worte: Die gesprochenen Worte weichen von jenen in Übungsdaten ab. Das kommt vor, wenn Leute einen starken Dialekt haben, eine veraltete Sprachform sprechen oder keine Muttersprachler sind. Beim Bearbeiten der Radio-Oranje-Sammlung tauchte dieses Problem auf, weil Königin Wilhelmina in den 1940er-Jahren eine Variante des Niederländischen sprach, die am Ende des 19. Jahrhunderts Standard gewesen war. Aufgrund von Änderungen der Recht-

schreibung erstellte der G2P-Konvertierer für modernes Niederländisch zudem fehlerhafte Transkriptionen des im Zweiten Weltkrieg verfassten Textes. Zum Beispiel wurde mensheid (Menschheit) als menscheid geschrieben, was zu einer falschen Phonemfolge führte: mEnsXEit anstatt menshEit. Mit einigen zusätzlichen Transkriptionsregeln und durch Hinzufügen phonetischer Transkriptionen nicht mehr gebräuchlicher Wörter konnte das Problem gelöst werden.

3. Unterschied zwischen gesprochenem und geschriebenem Text: Die manuelle Transkription ist eine ausgearbeitete Fassung der tatsächlich gesprochenen Sprache, d. h. oft tendieren Transkriptoren dazu, die gesprochene Sprache zu verbessern. Anstatt die mündliche Sprache mit all ihren Redundanzen, umständlichen grammatikalischen Konstruktionen oder Fehlaussprachen aufzuschreiben, schreiben die Transkriptoren diese in gutes Deutsch (bzw. Niederländisch oder Englisch) um. Sie erfassen die Absichten des Redners, ignorieren jedoch seine bzw. ihre Art zu reden. Für das Textverständnis mag dies kein großes Problem darstellen, aber es kann zu Kopplungsfehlern führen.

2. Oral-History-Sammlungen mit segmentbezogener Indexierung

Wir stellen im Folgenden einige Beispiele für Oral-History-Sammlungen vor, bei denen Spracherkennung oder Angleichungstools angewandt wurden, um sie durchsuchbar zu machen.

2.1 Radio Oranje

Im Zweiten Weltkrieg hielt die niederländische Königin Wilhelmina im Londoner Exil über Radio Oranje Ansprachen an das niederländische Volk. Diese Reden wurden der Königin von Beamten in London geschrieben. Sowohl die 37 auf Wachsplatten aufgenommenen Aufnahmen als auch die Transkriptionen als Durchschrift auf Kohlepapier konnten konserviert werden. Beides wurde 2005 im Radio-Oranje-Projekt digitalisiert und mithilfe von Spracherkennungstechnologie durchsuchbar gemacht.

Für dieses Projekt wurde eine Online-(Such-)Maske entwickelt, mit deren Hilfe

Ein Screenshot der Suchmaske des Radio-Oranje-Projekts. Das Wort „Rotterdam" kommt dreimal in den 37 Reden vor.

Screenshot der interaktiven Abspielmaske der Radio-Oranje-Aufnahmen. Das gesprochene Wort ist unterstrichen; das ausgewählte Suchwort ist fett markiert (Rotterdam). Der graue Balken repräsentiert das komplette Interview; die weißen Striche geben die Grenzen zwischen den geschriebenen Sätzen an. Der blaue Balken darunter zeigt die Vergrößerung der Zeitleiste für eine Minute Redezeit. Das gesuchte Wort ist durch den dicken roten Streifen im blauen Balken dargestellt. Der von Königin Wilhelmina gesprochene Satz diente zum Durchsuchen der Fotodatenbank „Beeldbank WO2" (www.beeldbankwo2.nl), um dazu passende Fotos auszuwählen.

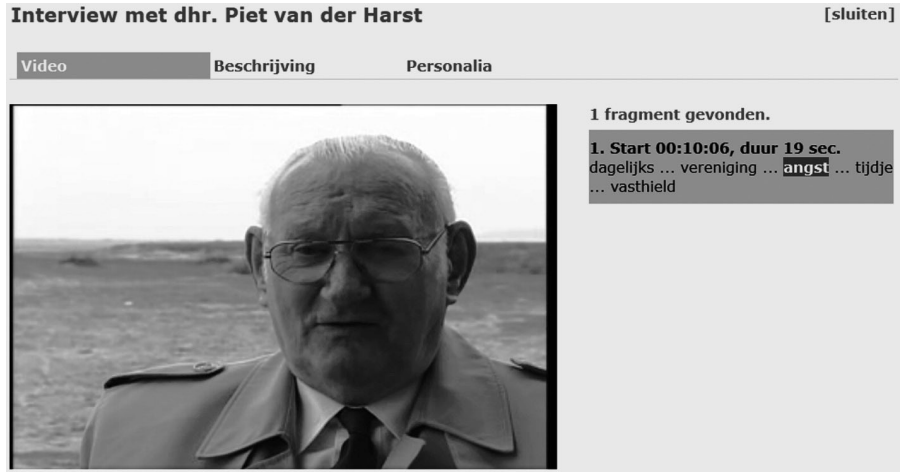
Nutzer die Transkription durchsuchen sowie die entsprechenden Audiofragmente abspielen und anhören können. Dies wurde durch automatische Kopplungstechnologie (Alignment) ermöglicht. Darüber hinaus werden zusammen mit der Audiodatei Untertitel angezeigt, wodurch die historischen, mitunter von Störgeräuschen übertönten Aufnahmen leichter zu verstehen sind. Das Projekt kann auf der Projektwebseite[5] der Universität Twente in den Niederlanden abgerufen werden.

5 http://hmi.ewi.utwente.nl/Showcases/Language-Multimedia-and-Information/Radio-Oranje-demo.

Startseite des Buchenwald-Portals (www.buchenwald.nl)

2.2 Buchenwald

Das Buchenwald-Portal bietet Zugang zu einer Sammlung von Interviews mit Überlebenden des Konzentrationslagers Buchenwald. In den 1990er-Jahren wurden 38 Interviews mit ehemaligen Gefangenen (alle männlichen Geschlechts) unter der Schirmherrschaft der „Niederländischen Vereinigung ehemaliger Buchenwaldhäftlinge" aufgenommen. Die Interviews samt dazugehöriger Textdaten (persönliche Profile und Zusammenfassungen) wurden mittels ASR segmentbezogen durchsuchbar gemacht. Ein zugehöriges Sprachmodell wurde entwickelt, das alle Arten von Text mit Lagerbezug einbezieht, um die von den Zeitzeugen verwendeten spezifischen Wörter zu erkennen.



Screenshot der Buchenwald-Webseite, durchsucht nach dem Wort „angst" (dt. „Angst"). Das Fragment, das (laut ASR-Maschine) dieses Wort enthält, beginnt bei Minute 10:06.

2.3 Lager Amersfoort

Zwischen 2000 und 2005 wurden 100 Interviews mit ehemaligen Insassen des von den deutschen Besatzern betriebenen „Polizeilichen Durchgangslagers" in Amersfoort[6] aufgenommen. Die ersten fünf bis zehn Minuten dieser 100 Interviews mit Zeitzeugen, die im Lager Amersfoort festgehalten worden waren, werden gegenwärtig transkribiert und dazu genutzt, um das akustische Modell eines jeden Sprechers

6 www.kampamersfoort.nl.

anzupassen. Ziel ist es, für das Lager Amersfoort ein spezifisches Sprachmodell zu entwickeln, damit die ASR-Maschinen das für Interviews dieser Sammlung spezifische Vokabular erkennen.

Schlussbetrachtung

Reine Spracherkennung bringt keine Resultate, die den Erfordernissen automatischer Untertitelung oder Transkription aufgenommener Interviews entspräche: Die Fehlerquote ist für diesen Zweck zu hoch. Ihr Potenzial liegt vielmehr darin, einen direkten Zugang zu Aufnahmen audiovisueller Sammlungen zu ermöglichen. Eine Transkription mit einer Fehlerquote von 40 Prozent liefert keinen lesbaren Text; doch können die richtig erkannten 60 Prozent der Wörter leicht über eine Volltextsuche abgefragt werden. So kann die automatische Spracherkennung zwar keine Untertitel erzeugen, wohl aber den Hörer bzw. Betrachter direkt zu jenem Fragment eines langen Interviews führen, in dem das von ihm gesuchte Wort fällt. Es ist anzunehmen, dass interdisziplinäre Kooperationen und das Training der ASR-Maschinen in naher Zukunft zu einer beachtlichen Verbesserung ihrer Leistung führen.