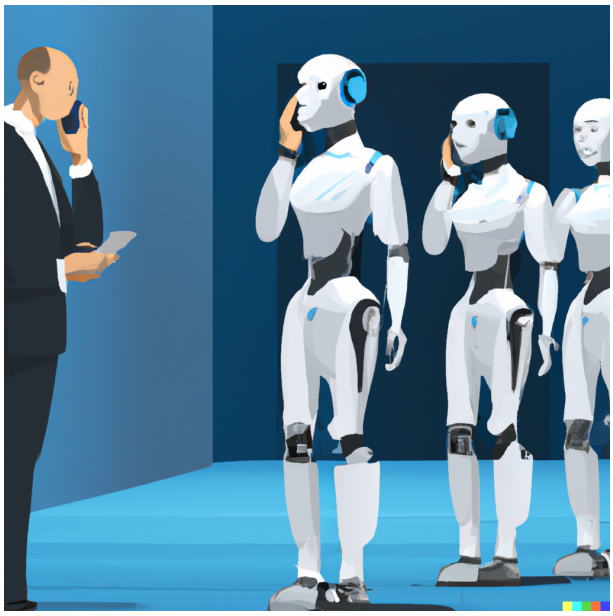
Text+ Task Area Collections

Jahrestagung

KI und Gesprochene Sprache



München, 27.-28.06.2024

Institut für Phonetik und Sprachverarbeitung
Bayerisches Archiv für Sprachsignale
LMU München
Schellingstr. 3
80799 München

www.phonetik.uni-muenchen.de
DOI: 10.5281/zenodo.12513374

Vorwort

Die Jahrestagung der Task Area *Collections* in Text+[1] zum Thema *KI und gesprochene Sprache* umfasst zwei eingeladene Vorträge aus der Industrie und einer neu eingerichteten KI Professur der LMU München sowie zehn Fachvorträge und einen Einführungsworkshop zu den Webdiensten des BAS.

Das Tagungsthema wurde bewusst weit gewählt – es reicht von der Forschung an KI Methoden bis zur Anwendung von KI Technologien in der wissenschaftlichen und industriellen Praxis, und bindet interdisziplinär verschiedene sprachverarbeitende Disziplinen ein.

Allen Vorträgen ist gemein, dass die modernen großen akustischen Modelle und die Sprachmodelle einerseits bisherigen Technologien deutlich überlegen sind, andererseits ganz neue Fragen aufwerfen. So ist z. B. die automatische Spracherkennung nun in bislang unerreichter Qualität auf dem eigenen Rechner verfügbar, so dass auch besonders schützenswerte Sprachdaten wie Patienten- oder Zeitzeugeninterviews in nicht kommerziell orientierten und sicheren Umgebungen zuverlässig transkribiert werden können. Allerdings liefert die automatische Spracherkennung (noch) nicht die für wissenschaftliche Arbeit notwendige Tiefe der Annotation, (noch) keine Erklärung, wie die Ergebnisse erzielt wurden, und (noch) keine befriedigende Antwort auf die Frage, welche Daten zum Trainieren der Modelle genutzt wurden oder unter welchen Bedingungen genutzt werden dürfen, und wie ein unerwünschter oder gar gefährlicher *Bias* vermieden werden kann.

Die Beiträge zur Jahrestagung zeigen eine Aufbruchstimmung. Die neue Technologie wird als große Herausforderung gesehen – im positiven Sinne, als Chance für neue Entwicklungen. Jeder Beitrag zeigt Neugier und Entdeckergeist, einige beschreiben bereits den praktischen Einsatz, andere eröffnen Perspektiven. Die nächsten großen Aufgaben werden sein, die KI in Richtung der wissenschaftlichen Fragestellungen weiterzuentwickeln und sie in universitäre Lehrpläne zu integrieren.

Ich freue mich auf die Tagung, und ich erwarte einen regen fachlichen Austausch und positive Impulse für die eigene Arbeit!

München, 27.06.2024
Christoph Draxler[2]

Danksagung

Ich bedanke mich bei Thorsten Trippel und Lukas Weimer vom Text+ Office für ihre Geduld und Unterstützung bei der Erstellung der Webseite, Organisation und Finanzierung, bei Camilla Anton und Eva Thoma für die tatkräftige Unterstützung, beim Congress Center der LMU für die Beratung und das Bereitstellen von Kaffeemaschine und Geschirr, und beim Lyrik-Kabinett München für die freundliche Aufnahme in seinen Räumen.

[1] Text+ (textplus.org) wird gefördert von der Deutschen Forschungsgemeinschaft (DFG) unter der Fördernummer 460033370

[2] Die Abbildung auf der Titelseite wurde generiert von openAIs generativer KI DALL·E v2 mit dem Prompt '3 humanoid robots talking via mobile phones with a human interpreter in the background'

Johann Prenninger

BMW Group

Aktuelle Herausforderungen in der Umsetzung von Machine Learning und KI Anwendungen in der BMW Group

Der Vortrag gibt einen Überblick über aktuelle Herausforderungen und die Umsetzung von KI-Initiativen bei BMW. Entlang des "Information Value Loops" werden die technisch-wissenschaftlichen aber auch organisatorischen Erfolgsfaktoren in den Bereichen Datenentstehung und Rückführung, BigData & Cloud, Advanced Analytics, Machine- und Deep Learning skizziert. Die zentrale Frage, die dabei beispielhaft beantwortet wird, lautet: Wie schaffen wir es, trotz zunehmender operativer Hindernisse im globalen Kontext mehr mit den Daten zu leisten, die Kunden besser zu verstehen und mit coolen on- und off-board Features effektiver im Alltag zu unterstützen? Analytics und Daten bilden dabei die elementaren Grundlagen, mit deren Hilfe KI ermöglicht und nachhaltig umgesetzt werden kann. Eine wichtige Basisdisziplin ist dabei auch das möglichst tiefgreifende Verständnis dessen, was neueste Technologien wie z. B. Generative AI und Large Language Modelle (LLM) zu leisten imstande sind.

Barbara Plank

AI and Computational Linguistics, Munich AI and Natural Language Processing

Natural Language Processing for Non-standard Language Varieties

Natural Language Processing (NLP) has so far largely focused on standard languages with many speakers and abundant data. Together with advances in architectures and computing, this has enabled the success of Large Language Models (LLMs). However, there is a significant asymmetry when it comes to small, non-standard languages. These language varieties have few speakers and/or no generally accepted standard, posing many challenges to LLMs and NLP due to scarce data and high variability. This is in stark contrast to large language models trained on massive amounts of data. In this talk, I will reflect on this asymmetry and how to process small languages with LLMs. I will also discuss the current challenges facing NLP research. These challenges include three major dimensions: lack of resources, modeling non-standard data, and human-centric design.

Inhhaltsverzeichnis – Contents

TAKE THE ATRAIN. INTRODUCING AN INTERFACE FOR THE ACCESSIBLE TRANSCRIPTION OF INTERVIEWS

Armin Haberl[a], Jürgen Fleiß[a], Dominik Kowald[b], Stefan Thalmann[a]

[a]Business Analytics and Data Science-Center, University of Graz
[b]Know Center & Graz University of Technology

The analysis of qualitative research data, like interviews and focus groups, is becoming increasingly important in various behavioral sciences. The analysis of such free form communication requires a transcript of the spoken word [1] and the transcription of audio data has long been a considerable cost and time factor. While the transcription of an interview of one hour requires up to six hours of manual work [2], advances in AI-based tools allow to speed up this process, significantly reducing the necessary transcription work to a fraction compared to manual transcription. However, so far using those AI tools necessitated using command line interfaces and installation processes that may pose a barrier to many potential users. aTrain, an open-source, offline transcription tool with a graphical interface (fig. 1), provides an easily accessible AI-based transcription of audio data in multiple languages. It requires no programming skills, runs on most computers, operates without internet access, and supports data privacy by not uploading data to external servers.

aTrain leverages OpenAI's Whisper transcription models, known for their accuracy and robustness comparable to human transcribers [3] combined with speaker recognition capabilities. The transcribed output seamlessly integrates with popular qualitative data analysis software like MAXQDA and ATLAS.ti, streamlining the research workflow. Available on the Microsoft Store for easy installation, aTrain is designed for speed and efficiency. It transcribes audio files at 2-3 times the audio duration on mobile CPUs using the highest-accuracy Whisper transcription models. With an entry-level graphics card, this speed significantly improves to 30% of the audio duration.

The development of aTrain was motivated by the limitations of existing transcription tools. While open-source transcription models like Whisper offer accurate transcriptions, they often lack user-friendliness and require programming knowledge [4]. Paid subscription tools, on the other hand, may raise data privacy concerns due to their cloud-based nature, especially in the context of sensitive personal information and GDPR compliance [5].

aTrain overcomes these challenges by providing a free, open-source, offline alternative that prioritizes data privacy and ease of use. Its local installation and offline operation ensure that data remains on the user's machine, addressing privacy concerns associated with cloud-based tools. The intuitive graphical interface eliminates the need for programming skills, making it accessible to researchers from diverse backgrounds. By combining cutting-edge transcription models with user-friendly features and data privacy considerations, aTrain empowers researchers to efficiently analyze qualitative data from speech interactions (fig. 2). Its accessibility, speed, and integration with QDA software contribute to a streamlined research process, facilitating deeper insights into the field.

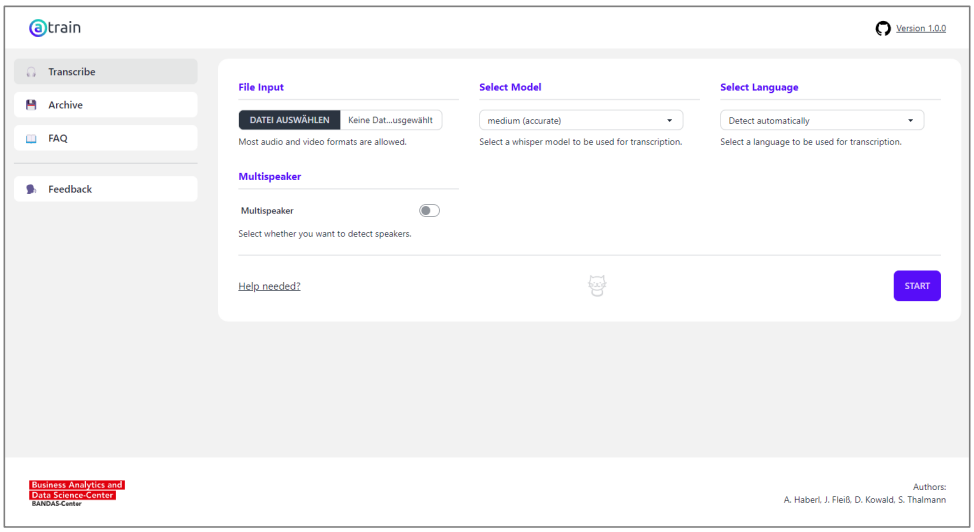


Figure 1 – : aTrain user interface

Figure 2 – : Transcription pipeline with corresponding inputs, tools and outputs

References

[1] BERSHADSKYY, D., L. DINGES, M. FIEDLER, A. AL-HAMADI, N. OSTERMAIERA, and J. WEIMANN: *Experimental economics for machine learning - a methodological contribution*. Tech. Rep., 2023. URL https://doi.org/10.24352/UB.OVGU-2023-104.

[2] BELL, E., A. BRYMAN, and B. HARLEY: *Business research methods (Sixth edition)*. Oxford University Press, 2022.

[3] RADFORD, A., J. KIM, T. XU, B. BROCKMAN, C. MCLEAVEY, and I. SUTSKEVER: *Robust Speech Recognition via Large-scale Weak Supervision*. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 28492–28518. 2023.

[4] WOLLIN-GIERING, S., M. HOFFMANN, J. HÖFTING, and C. VENTZKE: *Automatic transcription of qualitative interviews*. *Sociology of Science Discussion Papers*, 2023.

[5] *Regulation (EU) 2016/679 of the European Parliament and of the Council*. Tech. Rep., European Parliament Council of the European Union, 2016. URL https://data.europa.eu/eli/reg/2016/679/oj.

DEEP ASR - KÜNSTLICHE INTELLIGENZ IN DER SPRACHERKENNUNG: VERGLEICHENDE EVALUIERUNG FÜR GESPROCHENES DEUTSCH

Florian Schiel

Institut für Phonetik und Sprachverarbeitung, LMU München

Es gibt keine exakte Definition, was KI in der Spracherkennung (ASR) umfasst und was nicht. Heutzutage werden in der Praxis vorallem mehrschichtige neuronale Ansätze basierend auf sehr grossen Trainingsmaterialien, sog. Deep ASR, als KI bezeichnet (vor gerade mal 15 Jahren zählten auch statistische Modell wie GMM, HMM in allen ihren Varianten durchaus zur KI, vor 30 Jahren sogar regelbasierte Verfahren!).

Der erstaunliche Durchbruch der jahrzehntelang stagnierenden ASR mittels mehrschichtiger neuronaler Strukturen rechtfertigt diese Einschränkung; derzeit gibt es nur noch sehr wenige statistische Verfahren, welche Deep ASR Systemen Konkurrenz machen können (und meistens nur unter ganz bestimmten Einschränkungen). Daher macht es Sinn, sich mit der Performanz von aktuellen Deep ASR Systemen systematisch auseinanderzusetzen.

Ausformungen von Deep ASR sind vielfältig: verschiedene Architekturen, verschiedene Trainingsverfahren (*cost functions*), verschiedene Arten von Trainingsdaten (z.B. überwacht vs. unüberwacht), und nicht zuletzt die Tatsache, dass viele grosse Entwickler von Deep ASR keinen detailierten Einblick in ihre Methoden erlauben, macht es praktisch unmöglich, verschiedene existierende Systeme der Deep ASR quasi a-priori gemäß ihrer zu erwarteneden Performanz einzuordnen; in der Praxis bleibt daher nur der traditionelle vergleichende Performanztest mit definierten (und fairen) Metriken basierend auf einer unabhängigen (unveröffentlichten) Testdatenbank mit einem anerkannten Goldstandard (d.h. in der Regel einer kontrollierten Transkription des Gesprochenen).

Dieser Beitrag skizziert kurz, mit welcher Metrik und in welchen Testdatenstrukturen am Bayerischen Archiv für Sprachsignale (hdl.handle.net/11858/00-1779-0000-000C-DAAF-B) der LMU München in der Sprache Deutsch verschiedene verfügbare Deep ASR Systeme derzeit evaluiert werden. Der Vortrag strukturiert sich wie folgt: Beschreibung der Testdatenbank (der sog. *benchmark*) sowie ihrer inneren Struktur, welche es erlaubt, nach bestimmten definierten Aspekten von Sprache gesondert zu testen (in sog. *strata*), Definition der evaluierten Metrik Wortakuratheit, Auflistung der evaluierten ASR Systeme sowie einiger Besonderheiten von diesen, Probleme bei der Ausgestaltung einer fairen Evaluierung und deren Lösungen, und schließlich Diskussion der Evaluierungsergebnisse in verschiedenen strata und strata-Kombinationen. Zum Abschluss folgen praktische Hinweise, wie akademische Anwender die untersuchten ASR Systeme in den CLARIN BAS WebServices selber testen bzw. für ihre Forschungs- oder Lehrprojekte nutzen können (hdl.handle.net/11858/00-1779-0000-0028-421B-4).

Die wichtigsten Ergebnisse der Evaluierung sind: Deep ASR ist ein Quantensprung gegenüber statistischer ASR; der Umfang des Trainingsmaterials ist ein entscheidender Faktor; die Systeme mit den besten Ergebnissen tendieren dazu die Sprache zu glätten bzw. Inhalte zu erfinden, welche *top-down* sehr wahrscheinlich sind, aber nicht gesprochen wurden; strata Eigenschaften, welche im Trainingsmaterial nicht (oder zu wenig) repräsentiert sind, verschlechtern die Ergebnisse; *named entities* (z. B. Personen- oder Ortsnamen) sind ein Problem; fehlerfreie ASR gibt es nur für syntaktisch korrekte und fehlerfrei gelesene Sprache (z. B. Hörbücher, welche bedauerlicherweise oft als Trainings- und Test-Material Verwendung finden).

Warum die automatische Transkription noch nicht die Lösung aller Probleme ist

Jan Gorisch
Leibniz Institut für Deutsche Sprache

Ein lang anhaltendes Problem bei der Erschließung gesprochensprachlicher Aufnahmen ist der Transkriptionsflaschenhals. Bemühungen dieses Problem mit automatischen Methoden zu begegnen scheiterten in der Vergangenheit stets daran, dass die Erkennungsrate (oder dessen Pendant Wortfehlerrate - WER) in automatischen Spracherkennern (ASR) zu schlecht war. Systeme waren nicht auf die gegebenen Bedingungen (Akzent/Varietät, Hintergrundgeräusche, Echo, Aufnahmetechnik, Sprecher:innen-Anzahl, etc.) trainiert und konnten nur entsprechend schlechte Qualität zurückliefern. Vergleiche von Erschließungen mittels händischer Transkription vs. Korrektur durch ASR vortranskribiertem Text waren zwar vielversprechend [1], allerdings wurden hier die Daten vorselektiert aufgrund von Audio-Qualität und dem Grad an Spontanität. Das Aufkommen der neuesten Technologie, die mit tiefen Neuronalen Netzen und großen Datenmengen arbeitet, verspricht erneut den Transkriptionsflaschenhals beseitigen zu können.

Um dies anhand von Daten des AGD (Archiv für Gesprochenes Deutsch am Leibniz-Institut für Deutsche Sprache) im laufenden Betrieb zu prüfen, nahmen wir Aufnahmen (ca. 10h) verschiedener, teils historischer Korpora: Deutsch Heute (DH, 2006), deutsche Mundarten: Kreis Böblingen (BB) und Südwestdeutschland und Vorarlberg (SV, beide 1960er), und ließen zwei studentische Hilfskräfte nach Training an 45 Minuten Audio ihre konkrete Arbeitszeit notieren, während sie (i) von Null transkribierten, oder (ii) automatisch erstellte Transkripte korrigierten. Die Korrekturarbeiten beinhalteten neben der Korrektur des Inhalts auch das grobe Alignment von Text und Ton, sowie das Segmentieren maskierungswürdiger Stellen im Transkript, z.B. Nennung von Namen beteiligter Personen (Geburtsdaten und Orte), was bei der reinen Transkriptionsarbeit auch nebenher läuft. Des Weiteren sollten Äußerungen den entsprechenden Sprechern korrekt zugeordnet werden. Alle Arbeiten fanden im EXMARaLDA Partitur-Editor statt (exmaralda.org; [2]).

Für die Erstellung der automatischen Transkripte verwendeten wir OpenAI Whisper [3], zuerst das Modell *medium*; später (ab „order"-Nr. 16; s. Abb. 1) das Modell *large*. Die Sprecherseparierung erfolgte bei Aufnahmen mit separaten Kanälen je Sprecher:in entsprechend dem jeweiligen Energieanteil im Signal. Bei ein-kanaligen Aufnahmen verwendeten wir pyannote.audio [4], um die von Whisper transkribierten Wörter auf Sprecherspuren aufzuteilen. Orientieren sollten sich die Hilfskräfte an den Deutsch-Heute Konventionen, welche eine normalisierte Orthographie vorsehen.

Das Ergebnis war ernüchternd, vgl. Abb. 2: Hilfskräfte brauchten für beide Aufgaben den gleichen zeitlichen Aufwand von durchschnittlich 7,1 mal Echtzeit für die DH-Aufnahmen aus Hannover und Innsbruck. Aber ein signifikanter Effekt eines linearen Regressionsmodells war die Reihenfolge der Aufnahmen: je mehr Übung die Annotatoren hatten, desto schneller waren sie, vgl. Abbildung 2. Daraufhin ließen wir die Hilfskräfte nur noch ASR-Transkripte (nun basierend auf dem large-Modell) korrigieren, die wir für Aufnahmen aus Zürich (ebenfalls DH), sowie der historischen Korpora erstellten. Hier gab es neben der Reihenfolge auch weitere signifikante Effekte: die Annotatoren waren unterschiedlich schnell und der Einfluss der Korpora machte sich bemerkbar, vgl. Abb. 3.

Um im laufenden Betrieb die ersehnte Lösung ASR einsatzbereit zu bekommen, erwägen wir Möglichkeiten wie Fine-tuning, Prompting, oder weitere Einstellungen von z.B. Whisper per Kommandozeilenaufruf. Allerdings muss dann die Sprecherzuordnung aktuell weiterhin separat erfolgen, während sie in alltagstauglichen Systemen wie aTrain [5] bereits integriert ist. Auch das Wechseln zwischen Modellen, je nach Korpuseigenschaften, bleibt eine Option.

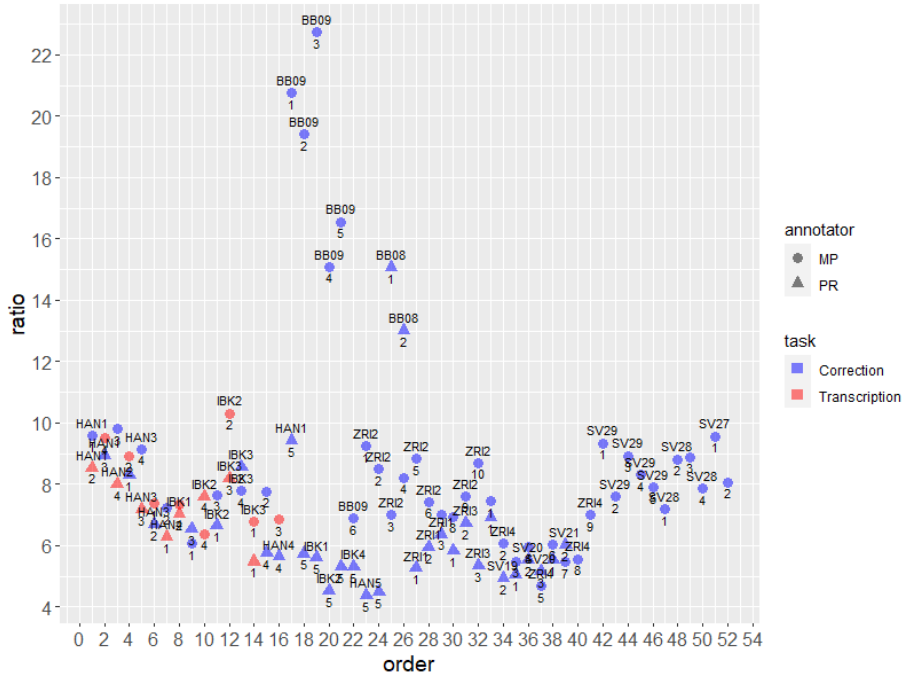


Abbildung 1 – Verhältnis (ratio) von Arbeitszeit und Audiolänge, gemäß der Reihenfolge (order) der Bearbeitung. Bis Aufn. 16: alternierend zwischen Aufgabe Transkribieren (rot) und Korrigieren (blau); später nur Korrigieren.

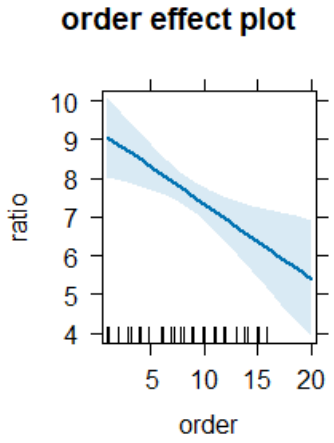


Abbildung 2 – Statistisch signifikante Effekte der Reihenfolge (order) über die ersten 16 Aufnahmen.

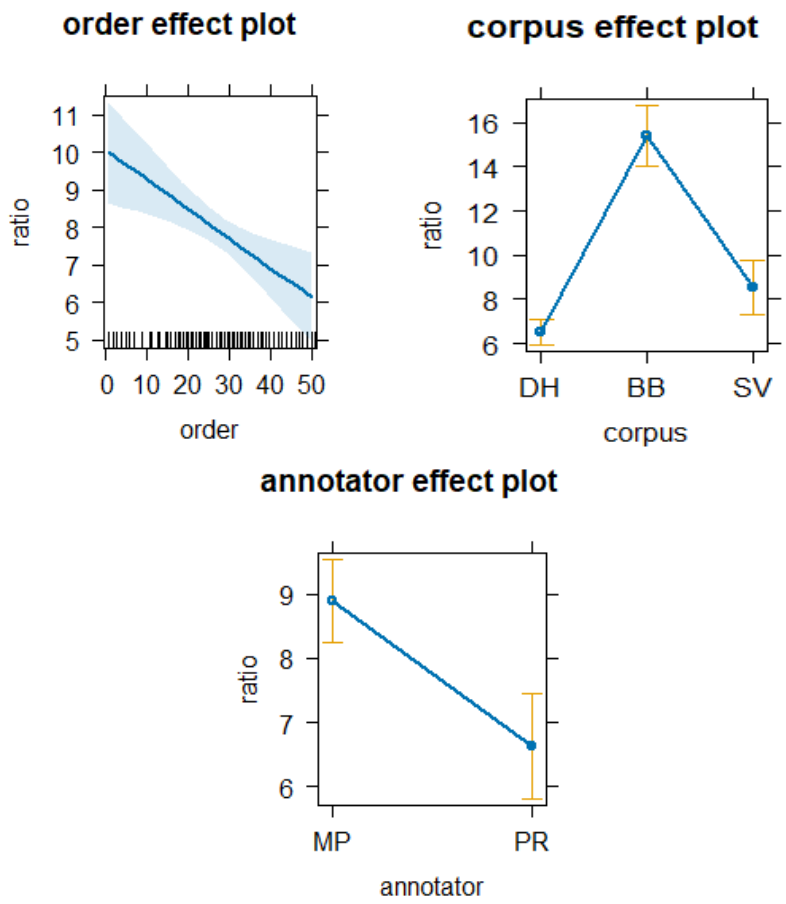


Abbildung 3 – Statistisch signifikante Effekte der Reihenfolge (order), Korpora (corpus), und Annotatoren über alle Korrektur-Daten.

Literatur

[1] BAZILLON, T., Y. ESTEVE, und D. LUZZATI: *Manual vs Assisted Transcription of Prepared and Spontaneous Speech*. In *Proc. Interspeech*. Brisbane, 2008.

[2] SCHMIDT, T.: *EXMARaLDA and the FOLK tools – two toolsets for transcribing and annotating spoken language*. In *Proc. LREC*. 2012.

[3] RADFORD, A., J. KIM, T. XU, B. BROCKMAN, C. MCLEAVEY, und I. SUTSKEVER: *Robust Speech Recognition via Large-scale Weak Supervision*. In *Proceedings of the 40th International Conference on Machine Learning*, S. 28492–28518. 2023.

[4] BREDIN, H.: *Pyannote.audio 2.1 Speaker Diarization Pipeline: Principle, Benchmark, and Recipe.* In *Proc. Interspeech*, S. 1983–1987. Dublin, 2023.

[5] HABERL, A., J. FLEISS, D. KOWALD, und S. THALMANN: *Take the aTrain. Introducing an Interface for the Accessible Transcription of Interviews*. *Journal of Behavioral and Experimental Finance*, 41, 100891, 2024.

Potential und Problemstellungen in der Automatisierten Bewertung der Kindlichen Sprach- und Sprechentwicklung

Hanna Ehlert, Lars Rumberg, Christopher Gebauer, Edith Beaulac, Maren Wallbaum, Ulrike Lüdtke, Jörn Ostermann

Auffälligkeiten in der Sprach- und Sprechentwicklung von Kindern umfassen die Rezeption und Produktion auf allen linguistischen Ebenen (Lexikon/Semantik: Wortschatz, Syntax/Morphologie: Grammatik, Phonetik/Phonologie: Aussprache und Pragmatik: Sprachgebrauch). Je nach Schweregrad können sie einen maßgeblichen Einfluss auf den weiteren Verlauf der emotional-sozialen, akademischen und kognitiven Entwicklung der betroffenen Kinder haben [1]. Deshalb ist die frühzeitige Identifikation solcher Auffälligkeiten mit dem Ziel den betroffenen Kindern passgenaue Intervention in Form von Sprachförderung oder Sprachtherapie zukommen zu lassen eine Querschnittsaufgabe des Gesundheits- und Bildungssektors (u. a. U-Untersuchungen, kontinuierliche Erfassung der kindlichen Sprachkompetenz in Kitas, Einschulungsuntersuchung). Für die Identifikation von Kindern mit Sprach- und Sprechauffälligkeiten kommen drei methodische Zugänge in Frage: strukturierte Beobachtung, systematische Spontansprachanalyse oder standardisierte Testverfahren.

In der Realität wird aufgrund von Fachkräftemangel, fehlenden zeitlichen Ressourcen oder mangelnder diagnostischer Ausbildung vor allem eine Methode vernachlässigt, die systematische Spontansprachanalyse. Spontansprachanalysen umfassen den Prozess der Transkription und linguistischen Analyse von in natürlichen Kommunikationskontexten aufgenommenen Sprachproben. Sie bieten gegenüber Testverfahren, welche Sprache in künstlichen Situationen stark limitiert erfassen, den entscheidenden Vorteil, dass mit ihnen die realen kommunikativen Fähigkeiten der Kinder auf allen linguistischen Ebenen bewertet werden können [2]. Trotzdem wird diese Methode in der Praxis wenig eingesetzt, da die Aufnahme solcher Sprachproben, deren händische, ggf. lautgetreue Transkription sowie die sich anschließende linguistische Auswertung extrem zeitaufwändig sind [3].

In unserem interdisziplinären Forschungsprojekt TALC (*Tools for Analyzing Language and Communication*) an der Stiftung Leibniz Universität Hannover erforschen wir deshalb den Einsatz von maschinellem Lernen (ML) in der Transkription und linguistischen Analyse von Kindersprache für diagnostische Zwecke. In diesem Beitrag werden vor allem die Herausforderungen beleuchtet, welche sich in der Entwicklung von Software zur automatischen Spracherkennung und -verarbeitung für diese Zwecke stellen. Sie ergeben sich aus der Art der Spontansprache (Menge, Variabilität, Kontext) im Zusammenspiel mit den jeweiligen Rahmenbedingungen der verschiedenen potentiellen Endnutzenden (u.a. pädagogische Fachkräfte, Sprachtherapeut:innen), den avisierten Analysemaßen und den Anforderungen an die Leistungsfähigkeit bzw. Fehlerrate der eingesetzten ML-Modelle.

Soll die Methode etwa in pädagogischen Kontexten wie Kitas zum Einsatz kommen, muss eine Durchführung der Erhebung der Spontansprache aufgrund zeitlicher und personeller Engpässe ohne Anwesenheit der Erzieher:innen gewährleistet werden. Dies senkt die Kontrollierbarkeit der entstehenden Sprachproben in Kombination mit einer deutlich anspruchsvolleren Trennung der Sprechenden (Extraktion der Anteile des Zielkindes aus der gesamten Aufnahme), was wiederum zu einer höheren Fehlerrate des Systems führen kann.

Unabhängig vom Elizitierungskontext sind neben der Sprecheridentifikation auch die nachfolgenen Prozessschritte, wie z. B. die automatische Spracherkennung, fehlerbehaftet und werden nie die Qualität einer manuellen Annotation erreichen. Analyseverfahren basierend auf ML können somit nur einen nachhaltigen Beitrag zur Identifikation von Kindern mit Sprach- und Sprechauffälligkeiten leisten, wenn sie systematische Fehler robust berücksichtigen und verlässlich eigenes Versagen detektieren können.

Literatur

[1] DODD, B.: *Differential Diagnosis of Pediatric Speech Sound Disorder. Current Developmental Disorders Reports*, 1(3), S. 189–196, 2014.

[2] HEILMANN, J. J., R. ROJAS, A. IGLESIAS, und J. F. MILLER: *Clinical Impact of Wordless Picture Storybooks on Bilingual Narrative Language Production: A Comparison of the 'Frog' Stories. International Journal of Language & Communication Disorders*, 51(3), S. 339–345, 2016.

[3] KLATTE, I., V. VAN HEUGTEN, R. ZWITSERLOOD, und E. GERRITS: *Language Sample Analysis in Clinical Practice: Speech-Language Pathologists' Barriers, Facilitators, and Needs. Language, Speech, and Hearing Services in Schools*, 53, S. 1–16, 2021.

COMPILING A PHONOLOGICALLY ANNOTATED CORPUS OF WEST AFRICAN ENGLISHES

Philipp Meer, Ulrike Gut
University of Münster

For the study of phonological properties of and differences between varieties of English, large speech corpora are necessary. Recent technological advances such as automatic speech recognition (ASR) tools have made their compilation much faster. In this talk, we report on the compilation of a corpus of Nigerian, Ghanaian and Cameroonian English for the study of vowel and consonantal differences between these varieties. First, we present how we enrich ICE Nigeria [1] with phonemic annotations, creating time-aligned phonemic transcriptions with the aligner (FAVE-align) of the bipartite Forced Alignment and Vowel Extraction program (FAVE; [2]), which has been adapted by Philipp Meer, in collaboration with a computer scientist at the University of Münster, for the analysis of Trinidadian English and other postcolonial varieties ([3], [4]). FAVE-align has been shown to perform well in the segmentation of different varieties of English compared with manual human alignment, even slightly more so than MAUS ([5], [6], [3]). Automatic alignment is followed by manual correction in Praat. Second, for Ghanaian and Cameroonian data we show how we create orthographic transcriptions using ASR and speaker diarization with WhisperX ([7]), followed by manual checks of correct speaker identification as well as utterance-level alignment and transcription accuracy in ELAN. As for the Nigerian English corpus, the time-aligned orthographic data will be subjected to phone-level annotation and segmentation using FAVE-align, followed by manual corrections in Praat. We will discuss the opportunities and challenges of using these tools for the compilation of phonological corpora.

References

[1] WUNDER, E.-M., H. VOORMANN, and U. GUT: *The ICE Nigeria corpus project: Creating an open, rich and accurate corpus. ICAME Journal*, 34, pp. 78–88, 2010.

[2] ROSENFELDER, I., J. FRUEHWALD, K. EVANINI, S. SEYFARTH, K. GORMAN, H. PRICHARD, and J. YUAN: *FAVE (Forced Alignment and Vowel Extraction). Program Suite v1.2.2*. 2014. URL `https://github.com/JoFrhwld/FAVE`.

[3] MEER, P.: *Automatic alignment for new englishes: Applying state-of-the-art aligners to trinidadian english. The Journal of the Acoustical Society of America*, 147(4), pp. 2283–2294, 2020.

[4] MEER, P., T. BRATO, and J. A. MATUTE FLORES: *Extending automatic vowel formant extraction to new englishes: A comparison of different methods. English World-Wide*, 42((1)), pp. 54–84, 2021.

[5] GONZALEZ, S., J. GRAMA, and C. E. TRAVIS: *Comparing the performance of forced aligners used in sociophonetic research. Linguistics Vanguard*, 6(1), pp. 1–13, 2020.

[6] MACKENZIE, L. and D. TURTON: *Assessing the accuracy of existing forced alignment software on varieties of british english. Linguistics Vanguard*, 6(1), pp. 1–14, 2020.

[7] BAIN, M., H. JAESUNG, H. TENGDA, and A. ZISSERMAN: *WhisperX: Time-Accurate Speech Transcription of Long-Form Audio*. In *Proc. Interspeech*. Dublin, 2023.

KI in der Korpusphonologie – Chancen und Schwierigkeiten am Beispiel des Korpus Tierras Altas y Bajas de Hispanoamérica (TiAlBA)

Verena Weiland

Universität Bonn

Der Vortrag stellt das Korpusprojekt TiAlBa (*Tierras Altas y Bajas de Hispanoamérica*) vor und diskutiert Herausforderungen sowie KI-gestützte Lösungsansätze in den Etappen

1. der Erstellung der orthographischen Transkriptionen,
2. der Erstellung der phonetisch-phonologischen Transkriptionen basierend auf IPA,
3. sowie der Datenauswertung.

Das Korpus TiAlBa umfasst derzeit Sprachaufnahmen von über 250 Personen aus 11 Ländern Hispanoamerikas. Es hat zum Ziel, erstmalig eine Datenbasis zur Analyse des gesprochenen Spanisch Hispanoamerikas im Bereich der Phonetik und Phonologie unter Berücksichtigung perzeptiver sowie akustischer Parameter zu schaffen. Die Absicht ist es, anhand dieser Daten die Gültigkeit der dialektalen Einteilung Hispanoamerikas in Hoch- und Tieflandgebiete (*tierras bajas/altas*, [1], [2]: 39) zu überprüfen. Diese geht davon aus, dass in den Küsten- bzw. Tieflandgebieten Konsonantenschwächung stattfindet (bspw. *gracias* > [ˈgɾasja]), bei den SprecherInnen aus den Hochlandgebieten Hispanoamerikas hingegen Vokalschwächung (bspw. *gracias* > [ˈgɾasjs]. Diese Differenzierung gilt in der hispanophonen Linguistik häufig als Faustregel, beruht jedoch auf impressionistisch dokumentierten Beobachtungen und entbehrt einer empirischen Datenbasis ([3]).

Eine wesentliche Herausforderung ist, dass die Sprachaufnahmen in einer natürlichen Umgebung anstatt in einem Phonetiklabor realisiert wurden. Die perzeptiven und akustischen Kriterien, die zu einheitlichen Codierungsweisen und zukünftig zur Vergleichbarkeit unterschiedlicher Studien führen sollen, sind eine wichtige Säule für die Codierung der Daten des Korpus TiAlBa. Softwarelösungen zur Transkription (*Praat, Fonométrica*), Auswertung und Visualisierung (*R-Statistik*), Dokumentation (*LaTeX*) und nachhaltigen Nutzung (universitätsinterne Datenbanken, *CLARIN-D*, *Text+*) sind weitere Bestandteile.

Die Sprachaufnahmen sind nach dem in [4] entworfenen *Protokoll (Inter-)Fonología del Español Contemporáneo ((I)FEC)* konzipiert und ermöglichen eine korpusphonologische Untersuchung des Spanischen. (I)FEC orientiert sich wiederum an den Forschungsprogrammen *Phonologie du Français Contemporain* (PFC, [5, 6], www.projet-pfc.net) und *Phonology of Contemporary English* (PAC, [7], www.pacprogramme.net). Zugrunde liegt diesen Forschungsdesigns die Untergliederung in mehrere Stufen der Autokontrolle im Prozess der Sprachproduktion nach [8] (208: „the amount of attention paid to speech“). Jede/r SprecherIn liest eine Liste von ca. 120 Wörtern sowie einen Text vor und gibt ein Interview anhand von Leitfragen. Die Lektüre von Wortliste und Text erlauben eine maximale Vergleichbarkeit der Aussprachrealisierungen aller SprecherInnen, während die Interviews gesprochene Sprache mit einem möglichst geringen Grad der Autokontrolle abbilden.

Literatur

[1] HENRÍQUEZ UREÑA, P.: *Observaciones sobre el español de América*. *Revista de Filología Española*, 8, S. 357–390, 1921.

[2] ROSENBLAT, A.: *El castellano de España y el castellano de América: unidad y diferenciaión (Cuadernos Taurus)*. Ed. Taurus, Madrid, 1973.

[3] DE CRIGNIS, P.: *Vokalschwächung im peruanischen Spanisch*. Ph.D. thesis, LMU, München, 2018.

[4] PUSTKA, E., C. GABRIEL, und T. MEISENBURG: *Romance Corpus Phonology: from (Inter-) Phonologie du Français Contemporain (I)PFC to (Inter-)Fonología del Español Contemporáneo (I)FEC*. In *Christoph Draxler/Kleber, Felicitas (Hrsg.): Tagungsband der 12. Tagung Phonetik und Phonologie im deutschsprachigen Raum*, S. 151–154. Munich: LMU, 2016.

[5] DURAND, J., B. LAKS, und C. LYCHE: *La Phonologie du français contemporain. Usages, variétés et structure*, S. 93–106. Narr, Tübingen, 2002.

[6] DURAND, J., B. LAKS, und C. LYCHE: *Le projet PFC. Une source de données primaires structurées*, S. 19–61. Hermès, 2009.

[7] DURAND, J. und A. PRZEWOZNY: *La phonologie de l'anglais contemporain: usages, variétés et structure*. *Revue française de linguistique appliquée*, 17(1), S. 25–37, 2012.

[8] LABOV, W.: *Sociolinguistic Patterns*. Blackwell, Oxford, 1972.

Automatische Transkription von Podcastfolgen für korpuslinguistische Untersuchungen. Ein Projektbericht.

Veronika Sahlbach, Vasco Alexander Sahlbach

TU Dresden

Podcasts sind aus der Medienlandschaft nicht mehr wegzudenken. Knapp die Hälfte der Bevölkerung über 16 Jahren hören hin und wieder Podcasts [1]. Sie überzeugen durch ihre besondere Tiefe und Ausführlichkeit, in denen verschiedene größere und kleinere Themen des Lebens besprochen werden. Diese reichen von Nachrichten, Hintergründen aus Wirtschaft, Politik, Wissenschaft und Gesundheit bis zu lebensweltlichen und komödiantischen Inhalten (vgl. [2], S. 2f.). Podcasts überzeugen durch die Auditivität, Immersion, Intimität, Engagement und eine starke Host-Bindung, wodurch das Format als eine Form des narrativen Journalismus vollen Erfolg erfährt (vgl. [3], S. 253).

Podcasts bieten sich für die Bearbeitung gesprächslinguistischer Fragestellungen an. Die Zuhörenden werden durch gekonntes Storytelling durch die Inhalte geführt, Situationen müssen aufgrund der fehlenden Bildunterstützung gut beschrieben werden. Doch nicht nur die Vermittlung der Inhalte erfolgt gekonnt, auch zwischenmenschliche Elemente werden durch die Performanz des Mündlichen übermittelt. Besonders in den unterhaltenden Formaten schlagen die Hosts oft einen umgangssprachlichen Plauderton an, wodurch eine freundschaftlich-nahe Beziehung zwischen ihnen und den Zuhörenden simuliert wird (vgl. [2], S. 258–263). Die Gespräche sind mehr als nur der Austausch von Informationen; sie erfüllen auch emotional-soziale Aufgaben. Bisher wurden intime Gespräche in Freundschaften oder Paarbeziehungen nur oberflächlich untersucht. Podcasts bieten hier eine neue Perspektive: Sprecher:innen diskutieren öffentlich zugänglich selbstgewählte Themen.

Diese emotionale Nähe soll in der angestrebten Erarbeitung sprachwissenschaftlich untersucht werden. Sprache entsteht in kommunikativen Welten, in denen geteiltes, transindividuelles Wissen ausgetauscht wird (vgl. [4], S. 15). Das Dissertationsprojekt hat zum Ziel, das situativ geteilte Wissen in sprachlichen Strukturen zu finden. Für die angestrebte korpuslinguistische Untersuchung wurden eine große Anzahl an Podcasts transkribiert und aufbereitet.

Die automatische Transkription größerer Datenmengen gesprochener Sprache stellte die Forschung jahrzehntelang vor zeitliche, technische oder finanzielle Herausforderungen. Obwohl die automatischen Transkriptionsgeschwindigkeiten in den letzten Jahren deutlich besser wurden, mussten Daten für linguistische Untersuchungen nachträglich korrigiert werden, um die Genauigkeit für eine Untersuchung zu erreichen. Mit der Veröffentlichung neuartiger Sprachmodelle wie Whisper wurde die Effektivität und Genauigkeit KI-basierter Transkriptionen erheblich gesteigert. Die mit Whisper vorgestellte robuste Spracherkennung basierend auf einem mit großem Korpus vortrainierten Sprachmodell erreicht sehr hohe Genauigkeiten (vgl. [5], S. 1). Das Modell wurde verwendet, um die große Datenmenge der zu untersuchenden Podcasts (ca. 450h Audiomaterial) zu transkribieren. Die Ausgabe des Sprachmodells umfasst eine reine Verbaltranskription. Für das angestrebte Ziel einer automatischen Minimaltranskription (vgl. [6], S. 359) ist eine Zuordnung der Sprechanteile zu den Personen notwendig. Dazu wird die KI-basierte Pipeline pyannote (vgl. [7]) eingesetzt, welche eine effiziente, automatische Zuordnung ermöglicht. Transkription und Sprecherzuordnung werden zu einem Minimaltranskript zusammengeführt und in einem Format kompatibel mit dem Partitureditor FOLKER abgelegt.

Ziel der Arbeit ist die Erprobung der verschiedenen zur Verfügung stehenden technischen Möglichkeiten sowie die linguistisch motivierte Verknüpfung für eine Dissertation zu Podcasts. In unserem Vortrag stellen wir die linguistische sowie technische Perspektive auf die Transkription gesprochener Sprache in großen Datenumfängen vor und erläutern das Vorgehen.

Literatur

[1] BITKOM: *Anteil der Befragten, die hin und wieder Podcasts hören, in Deutschland in ausgewählten Jahren von 2016 bis 2023*. 2023. URL https://de.statista.com/statistik/daten/studie/876487/umfrage/nutzung-von-podcasts-in-deutschland/.

[2] EINS, P.: *Podcasts im Journalismus. Eine Einführung für die Praxis*. Springer Fachmedien, Wiesbaden, 2022.

[3] MICHAEL, H.: *Podcasts als Erzählmedium. Grundmuster des Erzählens und die Intermedialität von journalistischem Storytelling in Podcasts*, S. 251–275. Springer Fachmedien, Wiesbaden, 2022. URL https://doi.org/10.1007/978-3-658-38712-9_10.

[4] HOFFMANN, L.: *Grammatik und gesprochene Sprache im Diskurs*, S. S. 5–28. De Gruyter, 2018. URL https://doi.org/10.1515/9783110538601-002.

[5] RADFORD, A., J. KIM, T. XU, B. BROCKMAN, C. MCLEAVEY, und I. SUTSKEVER: *Robust Speech Recognition via Large-Scale Weak Supervision*. 2022. URL https://doi.org/10.48550/arXiv.2212.04356.

[6] SELTING, M. ET AL.: *Gesprächsanalytisches Transkriptionssystem 2 (GAT 2). Gesprächsforschung*, 10, S. 353–402, 2009.

[7] BREDIN, H. und A. LAURENT: *End-to-end speaker segmentation for overlap-aware resegmentation*. 2021. URL http://arxiv.org/abs/2104.04045.

APPLYING EXPLAINABLE AI TECHNIQUES IN ACCENT RECOGNITION

Fritz Seebauer, Michael Kuhlmann, Reinhold Haeb-Umbach, Petra Wagner

Faculty of Linguistics and Literary Studies & CITEC, Bielefeld University, Paderborn University

For this workshop, we present an attempt to leverage current methods for explainable machine learning in order to gain insights into the acoustic properties of English accents spoken on the British Isles.

Finding out exactly which parts of a signal contributed to a perceptual distinction has proven elusive and is the topic of many branches of research in phonetics. Traditional approaches try to constrain the degrees of freedom of an experiment setting and try to search for correlation of carefully manipulated variables of potential acoustic or suprasegmental measurements. This is made necessary by the inability to simply query participants for their decision-making process as the answers might prove imprecise (see e.g. [1]).

To answer the question of which aspects of a given signal prove salient for distinguishing accents, we test a novel avenue of making a trained classifier highlight which parts of an input signal were most important for the final decision. We explore this paradigm on two different systems trained to classify English accents of a stratified subset from the Crowdsourced high-quality UK and Ireland English Dialect speech data set [2]. The first heuristic model is a modified version of the ACCDIST architecture proposed in [3]. The system classifies a given signal by comparing its articulation space denoted by MFCCs computed for each vowel against the pre-computed articulation spaces for each possible target accent. The model used in this investigation has been amended to use tri-phones instead of vowels as was suggested in [4] and the segmentation was obtained via a neural phonetic aligner trained on American English [5]. We extend this accent recognition model with an explanation module by computing the influence of each triphone, measured by its MFCCs contributions to the overall distance from the chosen accent. This generates a map of which triphones in the signal were most important in reaching the final decision. The validity of a generated explanation is also dependent on the accuracy of the classifier. Since the modified ACCDIST model performed rather poorly on the British Isles data, it was decided to train a second, more powerful classifier. The second model comprises a neural network leveraging two dimensional convolutions of log spectrograms to obtain embeddings which optimistically encode the properties of the accents under analysis. A posthoc explainer model as in [6] is then trained employing the speechbrain toolkit [7]. The explainer is based on vector quantization and transposed convolutions to obtain a decision mask regarding the parts of the original log spectrogram which were most salient for the classification.

Highlighting the regions of a signal which were important for a given classification decision allows for a more targeted phonetic post-hoc analysis. Computing acoustic features over a whole utterance and estimating their correlation to specific conditions potentially introduces a significant amount of noise. Consequently it would prove beneficial if it were possible to only regard those parts of the signal which are relevant for distinguishing accents.

We perform a perfunctory qualitative analysis of singular samples highlighting the advantages and pitfalls of this procedure on the accent dataset.

References

[1] KREIMAN, J. and B. R. GERRATT: *Validity of rating scale measures of voice quality*. *The Journal of the Acoustical Society of America*, 104(3), pp. 1598–1608, 1998.

[2] DEMIRSAHIN, I., O. KJARTANSSON, A. GUTKIN, and C. RIVERA: *Open-source multi-speaker corpora of the english accents in the british isles*. In *Proc. LREC*, pp. 6532–6541. 2020.

[3] HUCKVALE, M.: *ACCDIST: a metric for comparing speakers' accents*. In *International Conference on Spoken Language Processing, (ICSLP)*, pp. 29–32. 2004.

[4] HANANI, A., M. RUSSELL, and M. CAREY: *Human and computer recognition of regional accents and ethnic groups from British English speech*. *Computer Speech & Language*, 27(1), pp. 59–74, 2013. URL `https://doi.org/10.1016/j.csl.2012.01.003`. Special issue on Paralinguistics in Naturalistic Speech and Language.

[5] ZHU, J., C. ZHANG, and D. JURGENS: *Phone-to-audio alignment without text: A semi-supervised approach*. pp. 8167–8171. 2022.

[6] PAISSAN, F., C. SUBAKAN, and M. RAVANELLI: *Posthoc interpretation via quantization*. 2023. `2303.12659`.

[7] RAVANELLI, M., T. PARCOLLET, P. PLANTINGA, A. ROUHE, S. CORNELL, L. LUGOSCH, C. SUBAKAN, N. DAWALATABAD, A. HEBA, J. ZHONG, J.-C. CHOU, S.-L. YEH, S.-W. FU, C.-F. LIAO, E. RASTORGUEVA, F. GRONDIN, W. ARIS, H. NA, Y. GAO, R. DE MORI, and Y. BENGIO: *SpeechBrain: A general-purpose speech toolkit*. 2021. `2106.04624`.

KI-gestützte Workflow-Optimierung in Datenerhebungsprojekten – ein Werkstattbe-richt

Alina Hemmer

Universität Hamburg

Die rasch fortschreitende Entwicklung von Künstlicher Intelligenz (KI) und natürlicher Sprachverarbeitung (NLP) bietet spannende Ansätze, Datenerhebungsprozesse im Bereich der linguistischen Forschung effizienter zu gestalten. In einem Werkstattbericht werden die Anwendungsmöglichkeiten und Potenziale von Large Language Models (LLMs) in verschiedenen Phasen der gesprochensprachlichen Datenaufbereitung und -verarbeitung betrachtet und reflektiert. Der Fokus liegt dabei auf der Unterstützung der linguistischen Transkription von Audio- oder Videodaten, einem traditionell sehr zeit- und ressourcenintensiven Prozess, durch OpenAIs Whisper-Modell [1]. Zusätzlich werden Nutzen und Einbindungsmöglichkeiten anderer LLMs in das Post-Processing und die Weiterverarbeitung vor dem Hintergrund projektabhängiger Fragestellungen beleuchtet. Ziel des Werkstattberichts ist es, die praktischen Potenziale und Herausforderungen KI-gestützter Transkription und Weiterverarbeitung für die Sprachdatenerhebung zu ermitteln und Einsatzszenarien in linguistischen und interdisziplinären Projekten zu evaluieren.

Durch den engen Zeitrahmen vieler Datenerhebungsprojekte für gesprochene Sprache ist eine inhaltliche oder sprachliche Vorauswertung notwendig, um beispielsweise geeignete Ausschnitte im Datenmaterial auszuwählen, Anpassungen für weitere Datenerhebungsphasen vorzunehmen oder einen Überblick über das Vorhandensein relevanter Themen und sprachlicher Phänomene zu gewinnen. Für ein zeiteffizientes Vorgehen ist es erforderlich, nicht ausschließlich auf Audio- oder Videodaten zurückgreifen zu können. Gleichzeit ist der zeitliche Aufwand für die Erstellung manueller Transkripte sehr hoch und eine vollständige Transkription innerhalb der ersten Projektphasen in der Regel nicht leistbar. LLMs scheinen einen vielversprechenden Ausweg zu bieten und können an verschiedenen Stellen in Projektworkflows eingebracht werden. Die Veröffentlichung von Whisper durch OpenAI, einem fortschrittlichen Modell zur automatischen Spracherkennung (ASR), als open source ermöglicht die lokale Verarbeitung sensibler Audio- und Videodaten, welches eine wesentliche Anforderung vieler Datenerhebungsprojekte ist.

Ein zentraler Bestandteil des Werkstattberichts ist eine linguistische Analyse der KI-generierten Transkriptionen und die Evaluation verschiedener Konfigurationsparameter zur Optimierung der Transkriptionsqualität, wobei mögliche Einflussfaktoren, wie eine Varianz in der Qualität der Audioaufnahmen, berücksichtigt und – wo möglich – rekonstruiert werden. Als weiterer wichtiger Bestandteil wird erstes Feedback aus konkreten Datenerhebungsprojekten zu möglichen Effizienzgewinnen im Vergleich zu vollständig manuellen Transkriptionsprozessen vorgestellt. Auf Basis dieser Auswertungen werden erste Erkenntnisse zu Aspekten wie Zeitersparnis, einer Reduktion des Arbeitsaufwandes und der Verbesserung des Datenqualität gewonnen.

Neben der KI-gestützten Transkription wird das Post-Processing bzw. die Weiterverarbeitung automatisch erzeugter Transkripte mithilfe von LLMs als weitere Möglichkeit der Workflow-Optimierung betrachtet. Dabei geht es weniger um das allgemeine Potenzial von LLMs wie GPT-4 [2] oder Mixtral [3], sondern vielmehr um die praktischen Nutzungsmöglichkeiten im

Kontext datenschutzrechtlicher Einschränkungen und sensiblen Dateninhalten, die nicht außerhalb der eigenen Infrastruktur verarbeitet werden dürfen.

Der Beitrag soll als Grundlage für eine weitergehende Exploration der Potenziale von LLMs als Werkzeuge zur Optimierung von Workflows in Datenerhebungsprojekten dienen und zu einer generellen Diskussion beitragen, wie KI-Anwendungen in der linguistischen Forschung eingesetzt werden und möglicherweise zu einer Effizienzsteigerung beitragen können.

Literatur

[1] RADFORD, A., J. KIM, T. XU, B. BROCKMAN, C. MCLEAVEY, und I. SUTSKEVER: *Robust Speech Recognition via Large-scale Weak Supervision*. In *Proceedings of the 40th International Conference on Machine Learning*, S. 28492–28518. 2023.

[2] JOSH, A. ET AL.: *GPT-4 Technical Report*. 2024. 2303.08774.

[3] JIANG, A. Q. ET AL.: *Mixtral of experts*. 2024. 2401.04088.

RECORDING, VISUALISATION AND CLASSIFICATION OF OPTOPALATOGRAPHIC ARTICULATORY DATA

João Vítor Possamai de Menezes, Arne-Lukas Fietkau, Tom Diener, Peter Birkholz

TU Dresden

The acquisition and visualisation of articulatory data are essential not only for research in experimental phonetics, but also for Silent Speech Interfaces (SSI), which have gained significant research interest lately due to its promising applications, e.g., speech restoration [1].

Motivated by these applications, a variety of software capable of handling articulatory data have been developed for different sensing modalities, e.g., ultrasound [2], optopalatrography (OPG) [3] and electromagnetic articulography (EMA) [4].

Compared to other articulatory sensing modalities, OPG is very robust against session and speaker variability [5], it provides direct measurements suitable for speech therapy [6] and is under continuous development [7].

The system proposed in this paper is composed of 1) an OPG system used to record upper vocal tract articulation, 2) a custom software for recording articulatory data, and 3) a custom software to perform command word recognition.

The OPG system used in this work is an improved version of OPG Model 6 presented in [7]. This version uses a VCSEL and a phototransistor as an optical sensor pair. Fifteen sensor pairs are placed on a flexible printed circuit board to detect lip (2 pairs) and tongue (13 pairs) movements. To create an interface between the OPG and the recording software, a portable control unit was developed, which allows recording of articulation at a sampling rate of 100 Hz.

The articulatory recording software is based on the open-source Articulatory Data Recorder [3]. We developed the software further, resulting in the new version Articulatory Data Recorder 2.0 (ADR2). Operating offline, ADR2 can be used to record large datasets, which in turn are used by the recognition software to train and validate classification models. Operating online, it can be used to record single words upon which the recognition software performs inference based on the trained classification models.

The recognition software allows the visualisation of data, the training of classification models and the inference of new data recorded live by ADR2. Currently the only implemented classification model employs dynamic time warping in combination with a pattern matching algorithm. The training of the model consists of building word models for each of the command words within the corpus, that is, one feature vector containing the relevant feature variation patterns for each word. The command word recognition achieves 98.38% accuracy on a single speaker 40-word corpus using 5-fold cross-validation.

The presented system successfully integrates articulatory data recording and classification based on two dedicated custom softwares and a custom OPG hardware. The modularity of the classification system allows not only the integration of various corpora and classification algorithms, but also of different speakers, who can easily record new training data with their own OPG pseudopalates and be able to operate the demonstrator.

References

[1] GONZALEZ-LOPEZ, J. A., A. GOMEZ-ALANIS, J. M. MARTÍN DOÑAS, J. L. P. CÓRDOBA, and A. M. GOMEZ: *Silent Speech Interfaces for speech restoration: A review*. *IEEE Access*, vol. 8, 2020.

[2] WRENCH, A.: *Articulate Assistant Advanced User Guide: Version 2.17, Articulate Instruments Ltd.* 2017.

[3] WILBRANDT, A., S. STONE, and P. BIRKHOLZ: *Articulatory Data Recorder: A Framework for Real-Time Articulatory Data Recording*. In *Proc. Interspeech*. Brno, 2021.

[4] BUECH, P., S. ROESSIG, L. PAGEL, D. MUECKE, and A. HERMES: *"ema2wav": doing articulation by praat*. In *Proc. Interspeech*. Incheon, 2022.

[5] STONE, S. and P. BIRKHOLZ: *Cross-speaker silent-speech command word recognition using electro-optical stomatography*. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Barcelona, 2020.

[6] WAGNER, C., L. STAPPENBECK, H. WENZEL, P. STEINER, B. LEHNERT, and P. BIRKHOLZ: *Evaluation of a non-personalized optopalatographic device for prospective use in functional post-stroke dysphagia therapy*. *IEEE Transactions on bio-medical engineering*, vol. 6(1), 2022.

[7] BIRKHOLZ, P., S. STONE, C. WAGNER, S. KÜRBIS, A. WILBRANDT, and M. BOSSHAMMER: *A review of palatographic measurement devices developed at the TU Dresden from 2011 to 2022*. In *Proc. 20th International Congress of Phonetic Sciences (ICPhS)*. Prague, 2023.