An Innovative Methodology Utilizing AI-based Automatic Speech Recognition for Transcribing Dutch Patient-Provider Consultation Recordings

Cristian Tejedor-García, Radboud University, cristian.tejedorgarcia@ru.nl, ORCID: 0000-0001-5395-0438

Henk van den Heuvel, Radboud University, ORCID: 0000-0003-2064-0630

Arjan van Hessen, University of Twente, ORCID: 0000-0001-8888-7774

Sandra van Dulmen, Nivel, ORCID: 0000-0002-1651-7544

Toine Pieters, Utrecht University, ORCID: 0000-0002-8156-8436

Keywords

Automatic Speech Recognition, domain-adaptation method, AI, language modeling, Dutch medical discourse, patient-provider consultation recordings, deep learning

Abstract

In this paper, we present an innovative methodology for adapting special-purpose automatic speech recognition (ASR) models designed for sensitive audio data domains. This domain-adaptation method is based on evidence from pilot experiments conducted under the three-year funded project called Homo Medicinalis (HoMed-https://homed.ruhosting.nl) under the *Platform Digitale Infrastructuur Social Sciences and Humanities* (PDI-SSH).

In HoMed, we implement a SSH research infrastructure with significant potential for automatically transcribing sensitive audio-visual (AV) recordings. Employing artificial intelligence (AI) and deep learning (DL) algorithms, we established an automatic speech recognition (ASR) infrastructure for transcribing privacy-sensitive patient-provider consultation recordings. These files consist of video recordings of a medical appointment or consultation between a healthcare provider (such as a doctor or nurse) and a patient. The proposed methodology fully complies with the European General Data Protection Regulation (GDPR). We followed cutting-edge approaches and best practices in designing and implementing data pipelines, facilitating the seamless flow and transformation of data across various formats, institutions, and platforms. The significance of incorporating specialized training data focused on terms and phrases related to medication use in ASR systems is underscored in the HoMed project, as it has been observed that such targeted training mitigates errors, addressing a critical aspect in enhancing the performance of ASR technology (Mani et al., 2020). The challenge, naturally, is to develop a domain-specific ASR engine that can be publicly shared while keeping sensitive audio data securely on the premises of the owner. We have chosen in favor of an AI-based ASR technology approach in which there exists compelling need to improve its performance when applied to data that is specific to an in-house scenario (Tejedor-García et al., 2021) where the data is too sensitive to be transferred outside the institute.

The three most frequently used open-source state-of-the-art (SOTA) ASR models for Dutch in the period 2021-2023, Kaldi_NL (Open Spraaktechnologie, 2024; Araiza-Illan et al., 2024) in 2021, wav2vec2.0 (Baevski et al., 2020) in 2022 and Whisper (Radford et al., 2023) in 2023, were adapted and tested with patient-provider consultations since essential jargon was not included in their vocabulary lexicon (Litman et al., 2018). These ASR models underwent training and fine-tuning using existing radio and TV data in the

first phase (Tejedor-García et al., 2022; van der Molen et al., 2022) and using highly sensitive audio-visual recordings of patient consultations at the Nivel Institute (Utrecht) in the last phase of HoMed, for which we obtained the corresponding signed informed consent. We delivered an ASR-based infrastructure tailor-made for patient-provider consultation recordings that helps to improve the generic results of open-source Dutch ASR models and a method on developing domain-specific ASR engines with models that can be publicly shared based on sensitive audio data that cannot leave the premises of the owner.

The sensitive nature of the HoMed project poses the specific challenge of the restricted playback of the recordings (only possible at Nivel's video library), preventing sharing the raw data among institutes and project members. In the initial phase of HoMed, we overcame this limitation by utilizing training data resembling our use case from an alternate source (Tejedor-García et al., 2022)—the publicly accessible online news bulletins known as *Medicijnjournaal* from the Dutch Institute for Rational Use of Medicine (IVM). In the project's last phase, we used the sensitive data of medical conversations from Nivel, in which both patient and healthcare provider had signed informed consent for the recordings and for using the recordings for secondary analyses. We focused on devising a privacy-compliant method for handling the sensitive data itself. This involved addressing the constrained software and hardware environment in Nivel's viewing room, along with various restrictions on the use of the facilities. A crucial aspect is determining how to securely employ the sensitive data (or its derivatives, such as sensitive acoustic features and personal information in text transcriptions) to train the language, acoustic and lexicon models of the proposed ASR engines in an external secure location to Nivel. Throughout this process, we developed a standard protocol version for orthographic transcription of both Dutch medicine news bulletins and patient-provider medical consultations.

The methodology for preparing data for training and testing ASR models is illustrated in Figure 1. The process begins with the careful selection of audio data and transcriptions by an expert at the video library at Nivel. In particular, the data is only accessible via one computer in a local network inside the video library, which only a very few authorized researchers have access to. The data set focuses on patient-provider consultations, ensuring a minimal acceptable sound quality. Subsequently, the chosen data undergoes transcription by four native speakers *in-situ* following the same standardized protocol, resulting in a 50-hour dataset of 14 different topics (see Figure 2). Following transcription, feature extraction is performed using Mel Frequency Cepstral Coefficients (MFCCs) (Ittichaichareon et al., 2012). The resulting feature files, alongside the manual transcriptions, are securely stored in an encrypted folder on a USB stick. Notably, all these steps are executed within the confines of the video library at Nivel, ensuring a controlled and monitored environment.

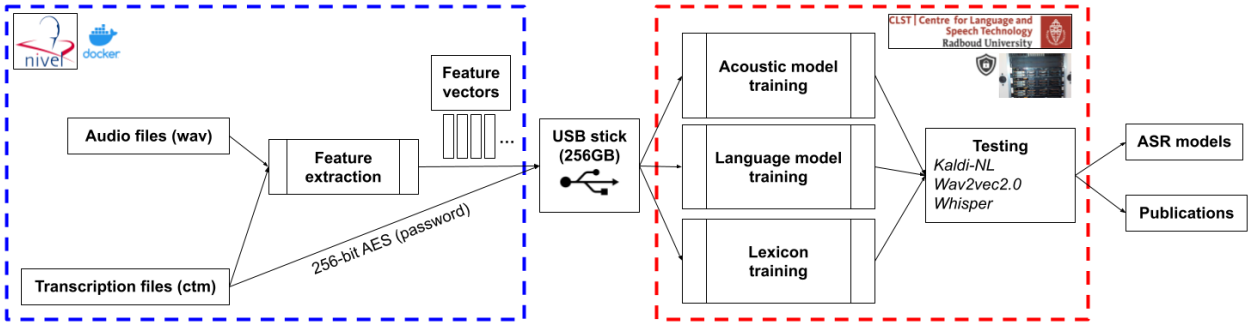

Figure 1. Implementation of an innovative methodology using AI-based ASR to transcribe Dutch patient-provider consultation recordings within the HoMed project.

The encrypted USB stick is then entrusted to an authorized individual, who transports it to a highly secured computer cluster for the training and testing phases of the ASR models, in this case at Radboud University. The training material contains 40 hours of speech of the total material transcribed in the first phase. Subsequently, the trained models are employed for testing on a 10-hour speech dataset with three different ASR systems, Kaldi_NL, wav2vec2.0 and Whisper. This robust methodology, executed at the NIVEL video library, demonstrates a comprehensive approach to data preparation for ASR models. Moreover, its adaptability makes it easily transferable to other domains requiring audio training, showcasing its versatility and applicability beyond the medical domain.

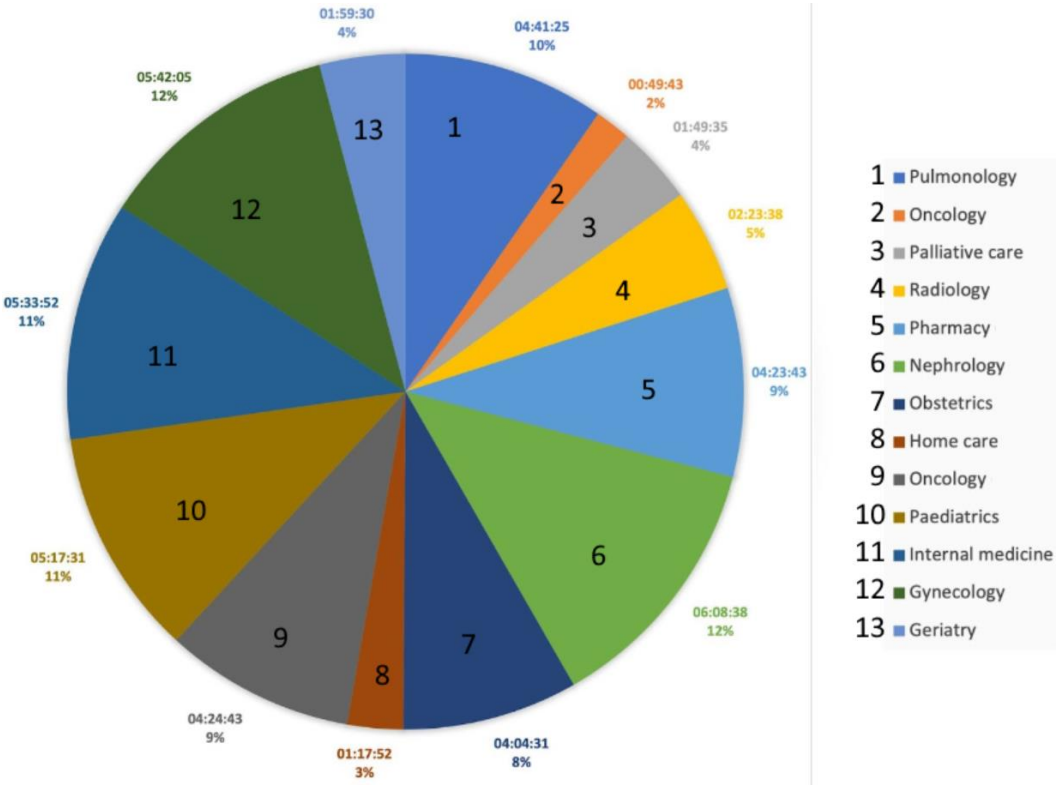

Figure 2. Transcribed material of patient-provider medical consultations, following the methodology proposed in HoMed.

In conclusion, this paper has introduced an innovative methodology for adapting special-purpose AI-based ASR models to handle sensitive audio data domain under the HoMed project. The project successfully implemented a research infrastructure for automatically transcribing privacy-sensitive patient-provider medical consultation recordings, utilizing AI and DL algorithms, complying with the GDPR and addressing the crucial need for accurate and secure transcription of medical consultations. By adapting and testing three SOTA ASR models for Dutch, the study has demonstrated the feasibility of improving generic results for open-source ASR models through domain-specific adaptation. The proposed methodology not only enhances the performance of ASR technology but also provides a blueprint for developing domain-specific ASR engines that can be publicly shared while ensuring the security of confidential audio information on the premises of the owner. This research contributes to the advancement of ASR technology in handling specialized audio data, particularly in the context of medical consultations, and holds promise for broader applications in sensitive audio-visual domains.

References

Araiza-Illan, G., Meyer, L., Truong, K. P., & Başkent, D. (2024). Automated Speech Audiometry: Can It Work Using Open-Source Pre-Trained Kaldi-NL Automatic Speech Recognition?. *Trends in hearing*, 28, 23312165241229057. https://doi.org/10.1177/23312165241229057 [accessed 10 March 2024]

Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33, 12449-12460. https://doi.org/10.48550/arXiv.2006.11477 [accessed 10 January 2024]

Ittichaichareon, C., Suksri, S., & Yingthawornsuk, T. (2012). Speech recognition using MFCC. In *International conference on computer graphics, simulation and modeling* (Vol. 9).

Kaldi_NL (Open Spraaktechnologie, 2024) <https://github.com/opensource-spraakherkenning-nl/Kaldi_NL> [accessed 10 January 2024]

Litman, D., Strik, H., & Lim, G. S. (2018). Speech Technologies and the Assessment of Second Language Speaking: Approaches, Challenges, and Opportunities. *Language Assessment Quarterly*, 15(3), 294–309. https://doi.org/10.1080/15434303.2018.1472265 [accessed 10 January 2024]

Mani, A., Palaskar, S., & Konam, S. (2020). Towards understanding ASR error correction for medical conversations. In *Proceedings of the first workshop on natural language processing for medical conversations* (pp. 7-11). https://aclanthology.org/2020.nlpmc-1.2.pdf [accessed 10 January 2024]

Radford, A., et al. (2023). Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning* (pp. 28492-28518). PMLR. https://doi.org/10.48550/arXiv.2212.04356 [accessed 10 January 2024]

Tejedor-García, C., Cardeñoso-Payo, V., & Escudero-Mancebo, D. (2021). Automatic Speech Recognition (ASR) Systems Applied to Pronunciation Assessment of L2 Spanish for Japanese Speakers. *Applied Sciences*, 11(15). https://doi.org/10.3390/app11156695 [accessed 10 January 2024]

Tejedor-García, C., van der Molen, B., van den Heuvel, H., van Hessen, A., & Pieters, T. (2022). Towards an Open-Source Dutch Speech Recognition System for the Healthcare Domain. In N. Calzolari, et al. (Eds.), Proceedings of the *Thirteenth Language Resources and Evaluation Conference, LREC* (pp. 1032–1039). ELRA. https://aclanthology.org/2022.lrec-1.110 [accessed 10 January 2024]

van der Molen, B., Tejedor-García, C., van den Heuvel, H., Ordelman, R., Pieters, T., van Dulmen, S., & van Hessen, A. (2022). Challenges on the Promising Road to Automatic Speech Recognition of Privacy-Sensitive Dutch Doctor-patient Consultation Recordings. *DH Benelux 2022 - ReMIX: Creation and alteration in DH (hybrid)*, Belval Campus, Esch-sur-Alzette, Luxembourg and online. https://doi.org/10.5281/zenodo.6517157 [accessed 10 January 2024]