Overcoming Data scarcity in the MediSpeech project

Authors:Cecilia Kuan, Radboud University
Maria Zamyrova, Radboud University
Khiet Truong, Radboud University
Maxim Popov, Amsterdam UMC
Abstract ID: 393
Event: Language Disorders 2026
Subject: 25. DELAD Special Session: Language Disorder Corpora
Presenter Name: Cecilia Kuan

Abstract

MediSpeech is a 3.5 years project running from 2025 to mid 2028 and is funded by ITEA4, see https://itea4.org/project/medispeech.html.
The project aims to reduce administrative waste in healthcare by creating an open digital healthcare ecosystem for automated medical reporting. The proposed solution uses AI-powered speech recognition, data interoperability and harmonisation, and technology-advanced clinical decision support to transform the healthcare model into a patient/doctor-centered approach.

One of the tasks is Speech-to-Text Conversion for patient-doctor consultations through Automatic Speech Recognition (ASR): This involves the real-time transcription of audio streams into text, specifically tailored for medical contexts. It emphasizes medical terminology to ensure accuracy and relevancy in clinical documentation. To do this, sufficient example recordings are needed to train, finetune and test ASR engines. Here, the project runs into data sparsity issues. Patient-doctor interactions are highly sensitive material which are hard to share even in projects where collaboration between partners is formalised. Given the current data scarcity, it is infeasible to train or benchmark such models effectively. This challenge affects not only individual researchers but also the wider Dutch NLP and clinical research community.

In this presentation, we report on methods found and explored on overcoming this problem. MediSpeech does not address speech of a pathological nature, but the data sparsity issues are similar for patient-doctor conversations. The methods explored to overcome this data sparsity issue are therefore relevant also for the DELAD community, and, thus, in our opinion, worth sharing.

We work on two paths to extend our data set: generation of synthetic dialogues and recording simulated patient-doctor recordings.

Generating high-quality synthetic data offers a practical solution to overcome these limitations by enabling effective model finetuning, robust evaluation, and system validation for anonymization, ontology mapping, and more. Synthetic data offers the advantage of being controllable, allowing targeted and balanced manipulation to introduce phenomena that are difficult to capture in real-world recordings, such as accented speech.

Finally, synthetic data can be shared freely, benefiting other researchers without the privacy constraints associated with real data. Using high-quality synthetic datasets can therefore aid the development of models and tools, a valuable alternative to real data for achieving desired performance levels.

The same holds for creating a set of simulated patient-doctor recordings. Recordings were made by Amsterdam UMC. The dataset currently contains 18 recordings of various lengths (5-45 minutes), totaling 4.5 hours. The transcriptions of these conversations will be manually corrected on the basis of a first transcription by an ASR. In this way we create a ground truth dataset for evaluation purposes.