

Beyond corpus building: Handling Multimodal Clinical Speech Data

Authors: Katarzyna Klessa, Adam Mickiewicz University in Poznań

Anita Lorenc, University of Warsaw

Łukasz Mik, AGH University of Krakow

Daniel Król, University of Applied Sciences Tarnow

Abstract ID: 391

Event: Language Disorders 2026

Subject: 25. DELAD Special Session: Language Disorder Corpora

Presenter Name: Katarzyna Klessa | Anita Lorenc

Abstract

Clinical speech corpora are becoming increasingly multimodal, combining speech recordings with instrumental measurements, clinical examinations and rich metadata. While much attention has been dedicated to corpus design and annotation, the management of multimodal clinical data throughout the research lifecycle still requires attention.

In this contribution, we report on experiences gained during the development of a multimodal corpus including recordings of disordered and typical speech. The project dataset integrates audio data with electromagnetic articulography (EMA), acoustic field distribution measurements for speech, CBCT imaging, and detailed clinical assessments conducted by speech therapists, orthodontists, and physiotherapists. These modalities differ in format, sensitivity and access requirements, yet all must remain linked through participant- and session-level metadata, making research data management an integral component of corpus development.

The corpus has been developed as part of an ongoing project investigating disordered speech and primary functions. Data are collected using a multimodal recording setup combining the CARSTENS AG501 articulograph, a custom-designed Acoustic Field Distribution Analyzer, and synchronized audio and video recordings. The participant group comprises individuals with functional articulation disorders, disorders of primary functions (e.g. tongue resting position and swallowing), and hypernasality or hyponasality, together with a control group of speakers with typical speech.

Drawing on this case study, we discuss the issues encountered and the solutions adopted throughout the development of the corpus. The case study highlights possible solutions for ethical approval, informed consent design for multimodal data collection, issues in managing sensitive personal and clinical information, and implementing strategies that separate identifying information from research data. We also discuss how file organisation and annotation workflows were designed to support secure storage, version control and future data sharing.

While the ethical approval permits the project-internal use of the complete dataset, public sharing is largely restricted to annotations and numerical data. A demonstration sample of the dataset has already been made available through the University of Warsaw Research Data Portal, and the remaining shareable data will be deposited after the completion of the project. Access to multimedia files remains substantially restricted in accordance with the ethics committee's decision.

Our experience highlights several lessons for managing sensitive multimodal research data. We acknowledge that approaches to ethical approval, data governance and data sharing differ across institutions and continue to evolve. Rather than proposing a universal model, we aim to contribute one practical perspective to the broader discussion on the responsible management and future reuse of multimodal clinical speech corpora, recognising the diversity of institutional practices and the continuing development of legal and technical frameworks.