



# Speech Annotation for A: Accuracy, Access, and Application

Zirong Li<sup>1</sup>, Hongchen Wu<sup>2</sup>, Yixin Gu<sup>2</sup>, Yao Du<sup>3</sup>, Yang Yue<sup>4</sup>

<sup>1</sup>Harris School of Public Policy and Department of Computer Science, University of Chicago, USA

<sup>2</sup>School of Modern Languages, Georgia Institute of Technology, USA

<sup>3</sup>University of Southern California, USA

<sup>4</sup>California State University San Marcos, USA

liz3@uchicago.edu, hwu480@gatech.edu, ygu321@gatech.edu, yaodu@usc.edu, yyue@csusm.edu

## Abstract

Accurate and efficient annotation of bilingual clinical recordings remains a persistent challenge, as existing solutions often require high demand for manual work by bilingual clinicians and their assistants and significant training related to annotation tools. To address this issue, we introduce Speech Annotation for A (SAFA)—an end-to-end, user-friendly “lazy mode” annotation workflow. By pairing annotation drafts generated from large language models with chunk-based editing, real-time difference highlighting, and speaker & language tagging - even in multi-speaker code-switching scenarios - SAFA delivers high-quality audio annotations ready for research with minimal setup and minimal human check. It further provides standardized CSV/TXT exports, bridging the gap between fully automated approaches and the meticulous accuracy demanded by multilingual clinical research, while facilitating the creation and expansion of high-quality labeled datasets for downstream studies.

**Index Terms:** bilingual clinical speech, end-to-end solution, research-ready annotation

## 1. Introduction

Annotating clinical recordings of bilingual data is notoriously time-consuming, often requiring hours for even short sessions [1]. When analyzing children’s speech and language data to determine underlying communication disorders, existing tools like ELAN (feature-rich but complex) and Praat (phonetics-focused) require considerable setup, while AI-driven systems such as SOAP.AI produce high-level summaries with limited control over time-aligned text [2]. To address these gaps, we introduce Speech Annotation for A (SAFA)—a user-friendly, end-to-end “lazy mode” solution that pairs large language model-generated transcripts with an intuitive interface, efficiently producing structured datasets ready for downstream speech analysis.

SAFA’s design aligns with InterSpeech 2025’s vision of inclusive speech technology—especially in multilingual environments in the healthcare context. SAFA embodies the principles of Accuracy (precise bilingual annotations), Access (zero-training interface), Application (clinical multilingual contexts), and Automation (LLM-driven transcript generation).

## 2. SAFA Workflow

SAFA annotation workflow is a unified solution for annotating audio and video data through a structured two-module approach.

Module One processes audio data using the automatic speech recognition (ASR) pipelines proposed and evaluated in [3]. Specifically, integrating speaker diarization (using PyAn-

notate) with OpenAI’s Whisper medium model produces the most accurate and well-performing annotation drafts. The output of Module One, also the input of Module Two, is a time-stamped annotation draft (in .json format) that includes speaker and language labels for each utterance.

This paper focuses on Module Two, a user-friendly annotation interface designed to minimize manual review while generating high-quality datasets for research. Module Two applies an MVC-inspired framework for developing applications using Python and PyQt6. The module maintains simplicity by moving some model tasks like transcript management and speaker identification into Controller and UI components but still keeps data and interface separation when feasible. VideoPlayer and TranscriptEditor PyQt6 widgets along with ChunkListWidget enable the View to deliver segment-based editing capabilities that sync with video playback. PyQt’s signal-slot mechanism enables the Controller to coordinate interactions that support synchronization and real-time annotation tracking.

## 3. UI/UX and main features

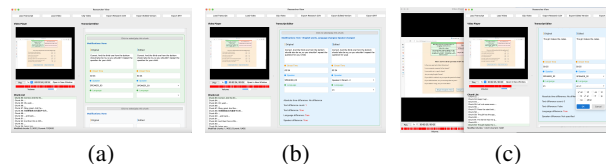


Figure 1: SAFA’s core UI elements in three sub-figures.

### 3.1. Chunk-Based Editing and Navigation

- Automatically divides LLM-generated transcripts into time-coded segments (Figure 1a).
- Synchronizes playback with the current segment, highlighting it automatically to streamline navigation.
- Clicking transcript chunks jumps playback directly to the relevant timestamp.
- Transcript view automatically scrolls with media playback for continuous visibility.

### 3.2. Real-Time Difference Highlighting

- Displays original versus edited text side-by-side, immediately highlighting edits with distinct colors (Figure 1b).
- Highlights changes to text, speaker tags, language tags, and timestamps instantly to ensure accuracy and consistency.

### 3.3. Speaker and Language Tagging

- Provides dropdown menus for quick selection of speaker roles (e.g., Clinician, Child) and languages (e.g., English,

Chinese, Spanish, with eight languages currently supported) (Figure 1c).

- Automatically calculates differences separately for multilingual segments, clearly distinguishing edits per language.

### 3.4. Video Clipping and Export

- Features an integrated “Clip” tool to extract specific video/audio segments without external software.
- Offers export functions producing structured, research-ready CSV, SRT and TXT files with detailed annotation.
- Plans to include additional export formats (e.g., JSON, VTT) to support broader subtitle and data analysis applications.

Start Time	End Time	Speaker	Language	Text
00:00:00	00:00:01	SPK001	zh	你好，我是王医生。
00:00:01	00:00:02	SPK002	en	Hi, I'm Dr. Wang.
00:00:02	00:00:03	SPK001	zh	今天感觉怎么样？
00:00:03	00:00:04	SPK002	en	Still a bit tired, but better.
00:00:04	00:00:05	SPK001	zh	好的，那我们继续吧。
00:00:05	00:00:06	SPK002	en	Okay, let's go on.

Figure 2: standard CSV export from SAFA

Figure 2 illustrates the standard CSV export from SAFA, clearly showing essential fields such as timestamps, speaker identities, language tags, and text revisions. This structured output format ensures broad applicability for diverse downstream tasks, including corpus-based research, linguistic speech analysis, machine learning model training, and natural language processing applications.

## 4. Comparison with Existing Tools



ELAN [4] provides robust multimodal annotation capabilities but presents a steep learning curve and uses tier-based configuration which tends to overwhelm non-technical users [1]. Praat [5] demonstrates exceptional capability in phonetic analysis through features like formants and pitch measurement but offers insufficient functionality for quick transcription of multiple speakers and typically necessitates custom scripts to handle metadata management. SOAP.AI [2] and Dragon Ambient eXperience (DAX) Copilot [6] produce high-level summaries of clinician-patient conversations using artificial intelligence but do not offer detailed control over time-aligned transcripts or the specific speaker and language labeling required for down-

stream language assessment, speech analysis, and natural language processing tasks.

SAFA’s intuitive, user-friendly “lazy mode” interface—combined with chunk-based editing, minimal setup, and real-time difference highlighting—supports detailed speaker and language tagging and enables seamless code-switching annotation. The system produces high-quality annotations that are immediately applicable to clinical, linguistic, and natural language processing research.

## 5. Conclusion

Speech Annotation for A (SAFA) provides an end-to-end solution for bilingual clinical annotation, guiding users from an LLM-generated transcript—delivered in a standardized, easy-to-parse format—through chunk-level validation, speaker and language labeling, and real-time difference tracking. Unlike fully automated pipelines that risk AI-induced errors, SAFA keeps the human in the loop, allowing non-technical clinicians to quickly validate and correct annotations with minimal setup.

By structuring metadata such as timestamps, speaker identities, and language tags—and automatically highlighting edits—SAFA reduces annotation time and ensures consistent data formatting. This not only enhances the quality of bilingual clinical reporting (e.g., by minimizing summarization errors from language models), but also lays the foundation for future multimodal analysis, integrating elements like parent-child speech, image references, or biometrics.

Ultimately, SAFA’s zero-training interface and flexible export options empower researchers to scale up bilingual annotated corpora, fueling advanced NLP and speech technologies and enabling more accurate, comprehensive evaluations in multilingual healthcare settings.

## 6. Appendix: Video

Video URL: [https://youtu.be/iqzEMo\\_wO6I](https://youtu.be/iqzEMo_wO6I)

## 7. References

- Z. Li, H. Wu, Q. Wang, B. Yao, D. Wang, R. Jia, and Y. Du, “Multimodal llms for children: Bilingual mandarin-english language assessment via telehealth,” in *2024 American Medical Informatics Association (AMIA) Annual Symposium, 2024*, poster presented at the AMIA Symposium, November 9-13, San Francisco, CA.
- Q. Zheng, P. Rabbani, Y. R. Lin, D. Mansour, and Y. Huang, “Soap. ai: A collaborative tool for documenting human behavior in videos through multimodal generative ai,” in *Companion Publication of the 2024 Conference on Computer-Supported Cooperative Work and Social Computing*, November 2024, pp. 87–90.
- H. Wu, Y. Du, Z. Li, Y. Gu, and D. T. Jayaprakash, “Evaluating automatic speech recognition pipelines for mandarin-english bilingual child language assessment in telehealth,” in *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech)*, August 2025, to appear.
- D. T. M. Caumo, M. P. França, and C. H. D. Silva, “Phonetic transcription of spontaneous children’s speech with the aid of software: A systematic review,” *Revista da ABRALIN*, vol. 21, no. 1, pp. 1–22, 2022.
- B. Gorjian, A. Hayati, and P. Pourkhoni, “Using praat software in teaching prosodic features to efl learners,” *Procedia - Social and Behavioral Sciences*, vol. 84, pp. 34–40, 2013.
- N. Communications, “Nuance Announces General Availability of DAX Copilot Embedded in Epic, Transforming Healthcare Experiences with Automated Clinical Documentation,” 2025. [Online]. Available: <https://www.nuance.com/healthcare/dragon-ai-clinical-solutions/dax-copilot.html>