High quality recordings and transcriptions of speech via remote platforms

Jonathan Wright, Robert Parker, Jeremy Zehr, Neville Ryant, Mark Liberman, Christopher Cieri, James Fiumara
Linguistic Data Consortium, University of Pennsylvania
Corresponding Author: Jonathan Wright jdwright@ldc.upenn.edu http://www.ldc.upenn.edu

Problem

- Numerous research domains leverage speech transcriptions; clinical diagnostic research in particular is an active area for LDC and others
- Commercial Speech to Text is now quite accurate but its deficits still leave room for using human transcription
- Notable deficits include omitting disfluencies (see bottom right of poster) and poor recognition in new vocabulary domains and unseen audio conditions
- While a commercial product may best target the intended utterance, much research requires accurate representation of the speech, including disfluencies
- However, the software landscape for recording and transcribing speech leaves much to be desired; no current platform to our knowledge combines all the features described here

Solution

- Custom mobile apps or web apps provide portable, remote, lossless recording options
- Web app allows remote and collaborative creation of transcripts
- Transcription interface has rich display and extensive keybindings
- Management interface allows coordination of large projects
- Transcription guidelines provide rich annotation and standard for training
- Cloud services provide extensible functionality
- LDC Speech Activity Detection provides pre-segmentation; humans correct boundaries
- Commercial Speech To Text inserted and then corrected saves time
- Validation of transcript in real time and second pass workflow for accuracy

Distributed Infrastructure

- Rather than a single tool, we use a distributed infrastructure of many parts
- GUI is a traditional web app with Ruby on Rails backend
- Web app runs within docker for portability, including local installation
- LDC platform is a combination of separate services on AWS
- web app, S3 storage, Speech Activity Detection, Speech to Text, transcript validation
- Web standards are increasingly powerful and browsers are increasingly standards compliant
- Javascript is widely used and has little danger of obsolescence

Transcription Interface

- Central to fast transcription speeds is a simple and intuitive interface
- At top is the left to right oriented waveform, at bottom is the top to bottom oriented transcript
- The tabular transcript is made up of timestamps, transcript, speaker label, and section label
- Blue lines below the waveform indicate the segments present in the transcript
- Red in the waveform and transcript calls out the selected segment
- A green waveform line and green segment background indicate playback position
- Navigation, playback, and labeling can be performed solely with the keyboard
- Numerous zoom, scroll, and playback options
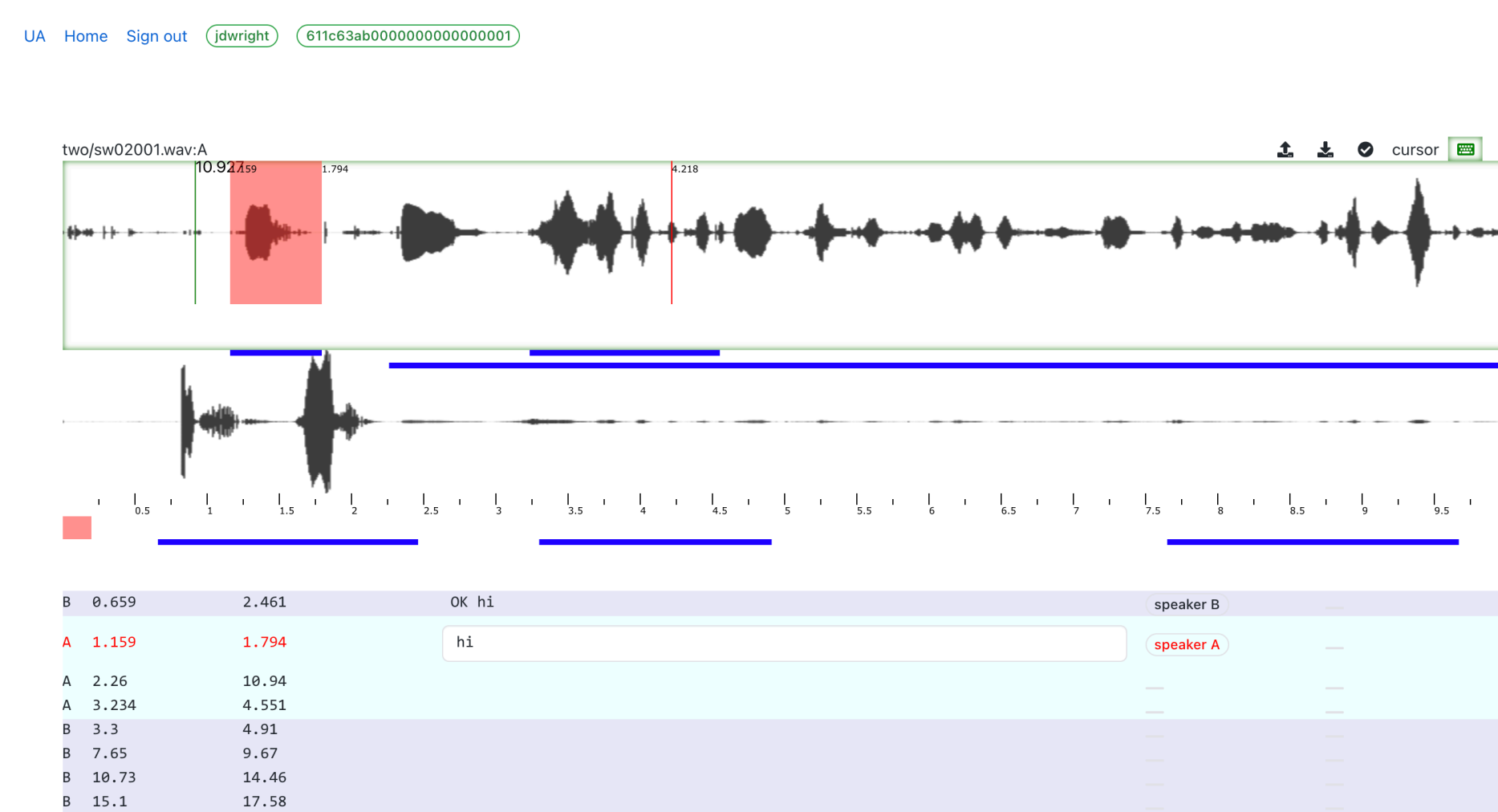
Channels, Speakers, Sections

- Channel separate from speaker
- Two channel mode automatic based on audio
- Speaker labels customizable and unlimited
- Labels could indicate some other dimension
- Section labels operate differently; they auto fill based on first and last segment

Management Interface

- A single instance can accommodate any number of users and tasks
- All users have accounts, both managers and transcriptionists
- Managers create tasks, add users to tasks, and create assignments
- Assignments can be followed from unassigned to assigned to done
- Percentage of dual assignments can be set
- First and second pass tasks can be connected; as assignments are completed in the first pass task, assignments in the second pass task are automatically created with transcript copies
- Transcripts can be downloaded as plain text, tab delimited tables

Special Transcript Elements

- filled pauses
- repetitions
- partial words
- speech errors
- non-speech noises
- unintelligible regions