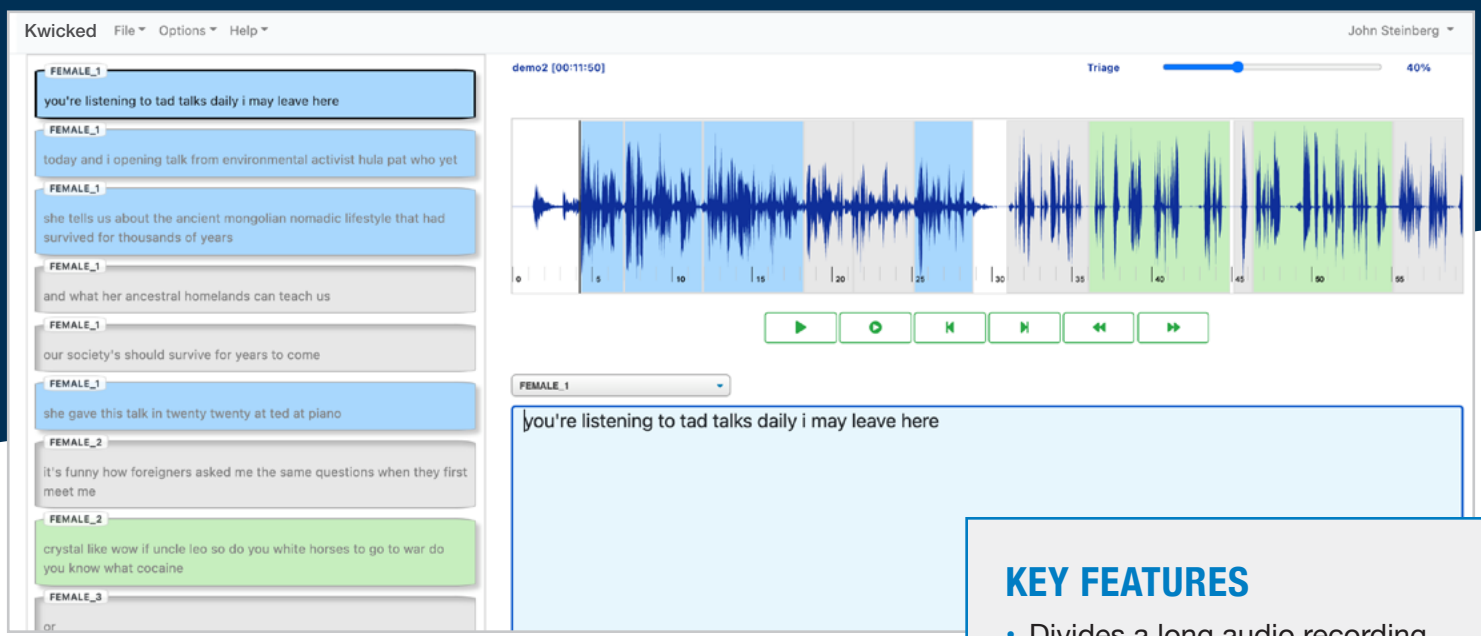
MIT LINCOLN LABORATORY

INNOVATION HIGHLIGHT

Technology in Support of National Security

WWW.LL.MIT.EDU

Kwicked Transcription and Captioning



With the Smart Transcription feature, users can correct the majority of errors created by an automatic speech recognition (ASR) system with a fraction of the work. Kwicked leverages confidence scores from the ASR system to recommend which parts of the transcription are more likely to have problems and require human correction. It's been shown that a majority of errors can be removed by correcting only 30% of the ASR transcript. Keyboard shortcuts allow users to skip inactive regions (marked with gray background) and only focus on correcting the active regions (green and blue backgrounds).

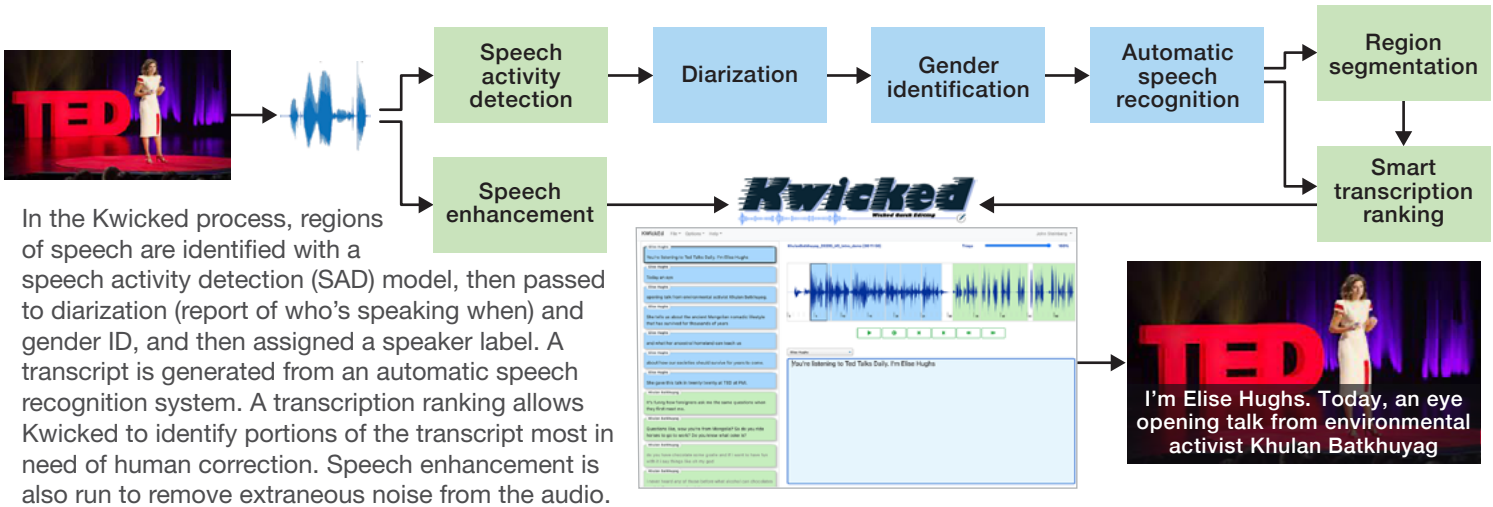
The growing demand for the accurate conversion of audio/video recordings to readable text prompted the development of Kwicked, a software tool that enables users to correct an automatically generated transcript or transcribe an audio message from scratch. Kwicked incorporates a "triage" approach to quickly identify errors in automatically transcribed speech so that human transcribers can efficiently make necessary corrections.

KEY FEATURES

- Divides a long audio recording into short segments to allow human analysts to avoid delays caused by pausing, rewinding, and fast forwarding the audio
- Provides a web interface that consolidates an audio player, a transcript browser, and a text editor to enable users to simultaneously control playback, transcript navigation, and transcription editing
- Uses speech enhancement to remove background noise, making the audio easier for human transcribers to understand
- Enables the creation of highly accurate transcripts at a fraction of the normal effort

LINCOLN LABORATORY
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

Kwicked Transcription and Captioning



In the Kwicked process, regions of speech are identified with a speech activity detection (SAD) model, then passed to diarization (report of who's speaking when) and gender ID, and then assigned a speaker label. A transcript is generated from an automatic speech recognition system. A transcription ranking allows Kwicked to identify portions of the transcript most in need of human correction. Speech enhancement is also run to remove extraneous noise from the audio.

Motivation

The growing reliance on video for disseminating news, tutorials, services, and entertainment has dramatically increased the demand for accurate transcriptions of speech, especially for the 11.5 million Americans with hearing loss. However, widespread access to accurate closed captions has been limited. Human-generated captions and transcripts are highly accurate, but their availability is cost-constrained by the salaries of skilled transcribers and costs of specialized hardware. While automatic speech recognition (ASR) systems are cost-effective solutions for large volumes of video/audio data in need of captioning, their transcripts often contain errors and differ significantly from what was spoken.

Lincoln Laboratory Approach

The Kwicked transcription tool offers a hybrid approach by which transcribers can either correct an automatically generated transcript or manually create a complete transcript of an audio passage.

Kwicked's advantage over commercial transcription software is its prioritization of portions of the ASR transcripts that are more likely to contain errors and need human correction. The Laboratory's approach strikes a balance between the ability of an automated system to process large volumes of recorded speech with the skill of a human to create a precise transcript.

Future Directions

- Open source the Kwicked algorithms to encourage both use and further development
- Tailor Kwicked for particular users and adapt the underlying ASR systems to acoustic and linguistic characteristics specific to a field, industry, or discipline
- Develop Kwicked for various languages, with a possible focus on uncommon languages for which there are few trained transcribers
- Enable Kwicked to use commercial speech engines so that users can choose whichever models work best for their needs

More Information

B. Borgstrom, M. Brandstein and R. Dunn, "Improving Statistical Model-Based Speech Enhancement with Deep Neural Networks," in *Proceedings of the 16th Annual International Workshop on Acoustic Signal Enhancement*, 5 Nov. 2018.

Approved for public release; distribution is unlimited. This material is based upon work supported by the United States Air Force under Air Force Contract No. FA8702-15-D-0001. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the United States Air Force.