Proposed Project Title: Hadoop Toolkit for Audio Analytics

Team Members: Kyle White

Current project proposal exists as described below while other project ideas are being considered, especially those that would fit into the Finance interest category.

Project Goals: The goal of this project is to develop the capabilities required to condition, decode, and understand speech content in an archive of audio data or across a number of audio streams. The conditioning portion of the toolkit is envisioned to only contain utterance segmentation and identification at the conclusion of the project, where utterance identification is meant to inform storage location since processing may be unique for each set of utterances. The recognition stage of the toolkit is currently envisioned to be comprised of support for running the Sphinx 4 recognizer on slave nodes, where the recognizer is configured for the locally stored speech data in terms of acoustic and language models. The speech understanding support stage is envisioned to be comprised of a finite state transducer to map input text to semantic key/value structures. Analysis will primarily be performed on the semantic key/value structures.

This problem calls for a big data solution on several fronts. Through parallelization a large number of audio streams can be analyzed, machine learning responsibilities distributed among nodes, large data volumes utilized since all data could be stored and algorithms re-run as any stage of the system changes.

Project Dataset: Controller/Pilot communications in the National Airspace. Data will nominally consist of controllers and pilots taking turns speaking on a single transmission frequency. The data set is available through my employer, and not public.

Language/Platform: JAVA, Python, MATLAB

Algorithm Descriptions: The utterance segmentation is planned to utilize GMM classification for initial VAD, and hypothesis testing via a criterion similar to BIC for determining where breaks occur between speakers. Utterance identification is planned to consist of identification of utterances as either controllers or pilots, and will be information already output from the segmentation pass that attempts to find breaks between speaker utterances. Work on this utterance segmentation algorithm has already been started from a project completed in E4810 at Columbia University. Recognition will be handled by Sphinx 4, with language and acoustic models being trained on a separate training set. It is planned that an open source toolkit such as openFST, or a Java FST implementation, can be used for semantic interpretation, where the configuration will need to be completed for the data of interest with the idea that the audio data will be stored in case the configuration changes and audio needs to be re-processed. Analysis will then be available on the semantic results, with the plan for analysis to be through existing Hadoop utilities to process the semantic key/value structures saved to a database on top of HDFS.