Sophia Davis
November 8, 2013
NLP Final Project Proposal

For my final project, my goal is to make a program that analyzes the probabilities of a spoken phrase being produced by various languages. I intend to combine the ideas of the n-gram/authorship project with further exploration of speech processing. This will probably involve adapting the textbook's algorithm for making a codebook of MFCC vectors for each time-window of training data, and incorporating these into some sort of language model.

For training and testing, I will use translated speeches from the proceedings of the European Parliament available at http://www.europarl.europa.eu/ (for example: http://www.europarl.europa.eu/http://www.europarl.europa.eu/sides/getVod.do?mode=chapter&language=DE&vodDateId=20120524-09:01:01-994). So far, I have figured out how to download the movies, but I am still working on transferring the files from mpeg4 video to wav format (I'm very sure I can get this working without too much effort).

I will be working alone on this project.