CSc 740 Project Proposal

1. Title: Sirí: A Spanish Accent Classification System for Voice Assistant ASR
   Names (alphabetical): Diego Freire, Martine Harrison, Stefanie Reed

2. This project addresses machine learning problems in ASR, specifically those concerning scalability, feature selection and engineering, and dimensionality reduction as they pertain to accents/dialects in speech models.

   The goal of the project is to achieve comparable word error rates for varieties of Spanish that are typically under-represented in training data, and to demonstrate the value of creating more robust ASR systems in preventing error propagation for downstream NLP tasks. Here, dialect refers to a particular form of a language, including its pronunciation, vocabulary, and grammar. Dialect identification in ASR is a special case of Language Recognition, where the system is required to discriminate between members of the same language, instead of different languages. Our dialect task is split into two sections: speech and text.

   <u>Speech Portion</u>

   - Create an acoustic model of Spanish speech data in multiple varieties by using a Hidden Markov Model-Gaussian Mixture Model that implements triphone modeling and speaker adaptation. After feature-processing, each phone is modeled without considering phonological context to create a monophone model, followed by building a triphone with the preceding and following phone of each target. The final speaker-adapted pass identifies which features make each phone maximally different through Linear Discriminant Analysis and Maximum Likelihood Linear Transform and by grouping the files by speaker for multiprocessing and cepstral mean and variance normalization.

   - Align transcribed utterances at the phone-level using XSAMPA forced-aligner.

   - Use a classification algorithm like SVM capable of identifying the Spanish dialects, using a dataset consisting of audio clips of spoken words (input features), and a transcript of what was spoken (target labels).

   - Compute accuracy using word error rate (WER), the de factor metric in ASR.

     WER = (Substitutions + Insertions + Deletions) / Number of Words Spoken
     Substitution: When a word is replaced (Shipping → Sipping)
     Insertion: When a word is added, and wasn't said (Hostess → Host is)
     Deletion: When a word is omitted from the transcript (get it done → get done)
     WER =25% is considered average.

   <u>Text Portion</u>

   - Identify and bin potential accents (according to corpus labeling scheme)

- Use pre-trained ASR on each dataset to see how SoTA out-of-the-box models perform on each accent, giving us a baseline. (E.g., Facebook's wav2vec performance on the overall Spanish Common Voice dataset has a 17.2% WER. But what would it be for Andean Spanish only, or Iberian Spanish only, etc.)

- Compiling polysemous words varying by region: Which words are most important for regionally-informed speech synthesis tasks? I.e. one region's term describing an everyday object could be considered derogatory language in another region.

- Processing text for the identified target words

- Dialogue Responses if we choose to do so

- Decide on a suitable embedding scheme: Word2Vec vs. FastText. FastText may potentially be better for highly inflectional Spanish morphology; i.e. FastText's embeddings for words with similar lemmas/stems might not be as distant from one another as Word2Vec's.

3. Team Member Roles
   Diego: Lead data visualization (Seaborn and Matplotlib), model selection, training and scoring (write WER and PER functions)

   Martine: Lead NLP (text data processing, word embeddings, dataset creation, data preparation and augmentation)

   Stefanie: Lead the speech processing. Data loading (SoX command-line tool, SciPy) and data cleansing (Scikit-Learn's feature selection module, for locating and excising constant, duplicate, or otherwise redundant features within the data), feature selection (mel-frequency cepstral coefficients for estimating formant values, principle component analysis for spectral and temporal acoustic features)

4. Related Topics each individual will contribute
   Diego: neural networks, model selection and training, performance measures
   Martine: text preprocessing, semantic analysis, lemmatization, feature generation,
   Stefanie: learning theory, dimensionality reduction, bias and machine ethics

5. Dataset description:

   The Mozilla Common Voice dataset is an open-source multi-language dataset of voices consisting of MP3 and corresponding text files. It is powered by the voices of people around the world who want to build voice applications. The project attempts to change the way machine learning models for voice applications are trained, where many of them are owned by companies, stifling innovation. Most voice datasets over-represent white and English-speaking males, limiting voice enabled technologies to work in just a few languages. Common Voice provides a diverse crowd-sourced alternative.

The dataset is subdivided into train, dev, and test portions with tags for high-quality data that is validated, and low-quality ones. The relevant metadata included is as follow:

- A speaker ID for each recording where each recording is ~ a sentence or less
- The path to the audio file
- A decoded audio array and its sampling rate
- A transcription of the sentence produced by the speaker
- How many upvotes the audio file received from reviewers
- Age
- Gender
- Accent based on region
  Accents include Central American, Andean-Pacific, Caribbean, Cliean, Central-Southern Iberian, Northern Iberian, Southern Iberian, Mexican, and Rioplatense.

6. Timeline

3/20	4/1 – 4/7	4/8 – 5/5	5/6 – 5/20
-Define topic and scope of project -Collect data	-Data preprocessing and exploration	-Define standards -Design, develop and evaluate model	-Documentation -Create video or poster -Adjustments

7. Demo

   In the time we are allotted to present our project, we hope to effectively describe the problem (ASR for Spanish dialects), as well as demonstrate our chosen algorithmic approach(es) alongside results.

8. Evaluation Plan

   - Standard ASR metrics (phone-error rate, word-error rate)
   - Optimize performance: Tune final model's parameters/hyperparameters before final comparison with wav2vec