Web-based Prosodic Labeling Toolchain

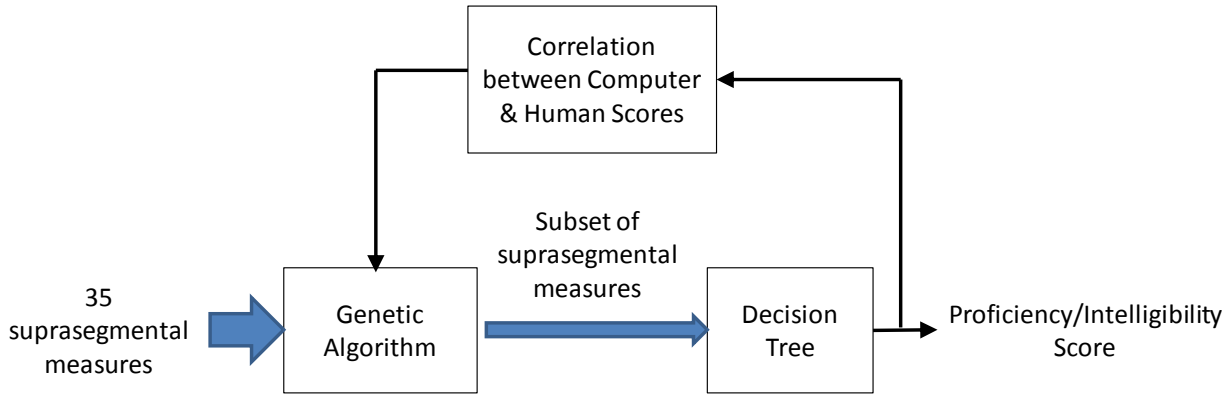
Sponsor Information:

Dr. Okim Kang
Department of English (TESL/Applied Linguistics)
Northern Arizona University
Okim.Kang@nau.edu

Project Description

In linguistics, prosody is the rhythm, stress, and intonation of speech. Prosody may reflect various features of the speaker or the utterance: the emotional state of the speaker; the form of the utterance (statement, question, or command); the presence of irony or sarcasm; emphasis, contrast, and focus; or other elements of language that may not be encoded by grammar or by choice of vocabulary. The Applied Linguistics Speech Lab (ALSL) developed a number of programs to extract prosodic features from audio speech files. These features are then provided as input to machine learning classifiers whose output is being utilized to measure the speaker's English proficiency and intelligibility. The current extracted prosodic features are based on a model of prosody developed by David Brazil. There is another popular prosody model based on the Tone and Break Indices (ToBI) system. The goal of this project is to extract the prosodic features based on the ToBI model and adapt the existing machine learning classifiers to measure proficiency and intelligibility based on the ToBI prosodic features. Then, the suitability of the two models for predicting proficiency and intelligibility can be compared by correlating the computer generated measurements with those assessed by humans. There are a number of freeware sources for code that automatically extract ToBI prosodic features that the Capstone team could use.

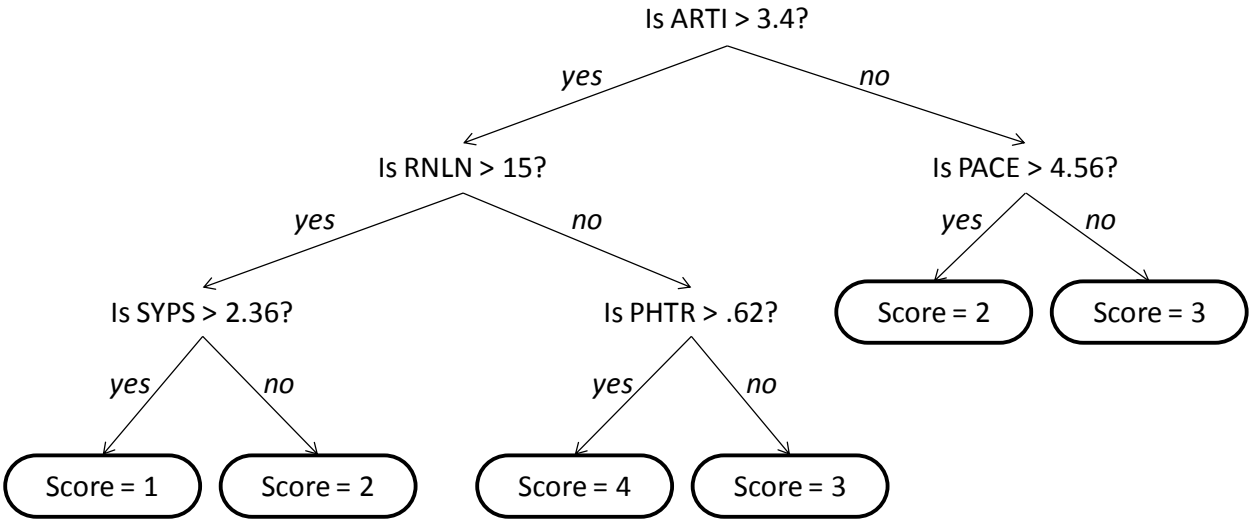
The current computer model is illustrated below:



First the computer analyzes the audio files and calculates 35 suprasegmental measures based on Brazil's prosody model. Suprasegmental measures include measures of how long the speaker pauses during the speech, what syllables the speaker accents, how many times the speaker uses "uh" or "um" and many more. Next a genetic algorithm selects a subset of the suprasegmental measures. A genetic algorithm is a problem-solving procedure employed by computers which is motivated by natural evolution, i.e.,

survival-of-the-fittest. In this case, the problem is to find a subset of suprasegmental measures which is the best (or fittest) at predicting the proficiency/intelligibility score. The procedure is repeated over-and-over with each repetition trying a different group of measures until the optimum solution is found.

During each repetition of the genetic algorithm, the computer develops a decision tree for predicting the proficiency/intelligibility score. A decision tree is a decision support tool that utilizes a tree-like graph of choices and their potential costs. It is depicted as a series of branching actions derived from evaluations of some numbers, in this case the values of the suprasegmental measures. The figure below is an illustrative example of a simple decision tree for predicting a proficiency/intelligibility score from the values of five suprasegmental measures (i.e., ARTI, RNLN, PACE, SYPS, and PHTR).



For each subset of suprasegmental measures that the computer employs to create a decision tree, it compares the scores generated by the decision tree with those assessed by humans with Pearson's correlation. This correlation is fed back to the genetic algorithm to guide it in selecting the next set of suprasegmental measures to evaluate. The end result of the process shown in the first figure is a set of suprasegmental measures and a corresponding decision tree that has the highest correlation between the computer's predicted scores and those appraised by humans.

The work for this project can be described as two tightly-linked parts: A prosodic labeler that extracts prosodic elements from audio files and analyses them; the second piece is to develop an easy-to-use web-based interface to this analysis tool.

Part 1: Prosodic Labeler. The goal of this part is to develop suprasegmental measures based on the ToBI model instead of ones based on Brazil's model. The team would select a freeware product that will extract the basic elements of the ToBI model from audio files. Then, the team will work with Dr. Okim Kang to develop code to convert these basic elements into suprasegmental measures. For example, one of the basic ToBI elements is the amount of accent, or stress, the speaker puts on syllables in an utterance. Brazil has only one level of stress (called prominence) while ToBI has multiple levels. There are several suprasegmental measures based on prominence, so these would all have to be converted to

an equivalent ToBI level of prominence. Once the ToBI-based suprasegmental measures are developed, the team would need to apply (and perhaps modify) the existing machine learning software (i.e., decision tree classifiers and genetic algorithm) to use the ToBI-based suprasegmental measures.

Part 2: A simple convenient web interface. The suite of independent software products currently used to do this work is divided between two operating systems, Windows and Linux. The LINUX software is an implementation of an open source automatic speech recognizer called KALDI, and processes the raw audio files to extract phones, which are the basic sounds which make up all human languages. The LINUX software is composed of shell scripts and C++ software which comes with the KALDI package. The rest of the software, which is Matlab-based, currently resides in a standard Windows operating system. This software takes the phones produced by KALDI and combines them into syllables (using a product developed by a previous Capstone team). The syllables are then analyzed by various machine learning tools (e.g., neural networks, decision trees, support vector machines) to create the 35 suprasegmental measures, which are then processed by more Matlab code described above to produce proficiency and intelligibility scores.

Obviously, the fact that this tool chain is spread across several independent software products running on several different platforms is very inconvenient and create lots of inefficient effort for our group. The goal of this second part of the project is to develop a simple web interface to tie all of these tools together. Users could create accounts in a web-based system, could upload their audio files, clink to run them through the tool chain (showing status along the way), and then receive not only the resulting proficiency and intelligibility scores for uploaded audio files, but also the suprasegmental measures utilized to predict these scores.

Knowledge, skills and expertise required for this project

- MATLAB, Java, and possibly utility languages like Python.
- Java

Equipment Requirements

- MATLAB
- Development environments for Java, etc.

Deliverables:

- A functioning prosodic element extraction and labeling tool, including a command-line interface. Installed on the ALSL
- A functioning web2.0 website outlined above, installed either on a local server or (preferred) in the free tier of cloud-based virtual solution like AWS.
- Installation of the tool chain described above within the new web-based site.
- Modification of existing Matlab software to use the ToBI model instead of the Brazil model to predict proficiency and intelligibility

- Regular progress meetings
- User manual, for non-technical end-users.
- Documentation sufficient that another programmer could add additional functionality to the code with minimal effort