Speech Recognition and Understanding Proposal
"Applause" Language Tutoring Application

Lara Martin, Vinay Vemuri, Nikolas Wolfe
Advisors: Florian Metze, Alan W Black, Maxine Eskenazi

Overview of the Problem

The Peace Corps is a development and cultural exchange volunteer program run by the United States government. Since 1961, more than 210,000 Americans have served in 139 countries around the world, working in such areas as social and economic development, education and public health. Volunteers typically commit to serve in a given country for 27 months, and in many cases become proficient (or better) in the languages of the communities in which they live and work. Unfortunately, this task is rendered exceedingly difficult at times by the lack of linguistic training resources available for many languages spoken and encountered by Volunteers. For obvious reasons, this creates a serious practical barrier to the efficacy and productivity of Volunteers in the field.

As of this writing, no robust platform currently exists for Peace Corps Language and Cultural Facilitators (LCFs) or Volunteers to create language training resources themselves. Independently, some Volunteers have initiated efforts such as the Celebrate Language Audio Project (CLAP) and Bantu Babel, but for a variety of reasons these projects have struggled to take root and become widely utilized. Although modern state-of-the-art language technologies have not been brought to bear in any of these initiatives (yet), the interesting juxtaposition of Volunteers and speakers of ultra-low resource languages presents a unique opportunity. The Peace Corps is in an arguably excellent position to simultaneously benefit and contribute to world heritage through the creation of language technologies for ultra-low resource languages. Broadly speaking, the purpose of this research project is to collaborate with the Peace Corps to build the necessary foundational tools to create quality language training materials for ultra-low resource world languages.

Proposed Solution

Often times, Peace Corps Volunteers are invited to their countries of service several months in advance. It would be useful to create high-quality language-learning materials to disseminate to Volunteers prior to Pre-Service Training (PST). Conventional learning tools are inapplicable in this situation because, almost by definition, the languages Peace Corps Volunteers speak are some of the lowest resource languages in the world.

Our initial plan is to build and evaluate an ultra-low resource language-independent pronunciation scoring tool to be used by the Peace Corps as part of a pre-departure online training program for new Volunteer invitees. The application will help improve Volunteers' pronunciation and recall for various introductory phrases in a target language. In essence, our goal is to find a way to measure and properly identify the "quality" of non-native speaker pronunciation. The set of phrases for which this tool will be used is "canned" and rather simple, and in the future we hope that this will allow for new modules in various target languages to be created rapidly. Much of the work done for the CLAP project is likely to be reused in this capacity. We currently have several hours of recorded, labeled, high-quality speech in twelve (12) low-resource languages from Ghana and Mali.

Our general approach will be to make some novel use of Mel-Cepstral Distortion (MCD) measures to get a sense of the "distance" between a non-native speaker's pronunciation of a given phrase and a target recording of a native speaker. This methodology has been employed by Alan Black et al to judge the quality of voice synthesizers by computing a difference between a synthesized voice and the target recording from which it was generated. While we expect that this is likely to have a smaller (better) MCD value than the comparison of a native speaker and a different person (with a conceivably different sounding voice), we postulate that the distance measure, on the whole, will decrease as a given non-native speaker gradually attains higher levels of proficiency. We intend to determine the correlation between objectively improving MCD scores and the subjective notion of "improvement" or "good pronunciation" by doing a study where native speakers of several test languages offer their assessment of non-native speakers of varying proficiencies pronouncing a set of predetermined phrases. For this study we intend to make use of the wide variety of native speakers of low-resource languages in the Language Technologies Institute as well as the LCFs from Peace Corps Ghana, several of whom have voluntarily offered their help with this project.

Implementation Details

With respect to implementation, the target phrases we will use (that is, the phrases recorded by native speakers) will be either be segmented or not segmented. In general we do not have recognizers for low-resource languages and so the question of how we will create such a segmentation bears examination. We propose that, to segment phrases in an arbitrary low-resource language, we will either:

1.) Choose one or more non-target language phoneme recognizer(s)
2.) Use a non-language-specific IPA feature detector for phones.

If we use multiple language recognizers, we would like to use recognizers for which there is as much phonetic diversity as possible so as to capture information about a larger range of detectable phonemes. However, our expectation is that Option 2 above will likely yield a more expedient segmentation, for the simple reason that it does not produce language-dependent segmentations.

As a way of aligning non-native speaker pronunciation with native speaker pronunciation, we propose to use Dynamic Time Warping (DTW) to align the the target phrase (whether segmented or not) with user-generated speech. We will then compute an MCD measure for each phone pair, and combine these individual scores in some novel yet-to-be-determined way. The MCD scores within the phones will either be averaged (so that the length of the user's speech is not taken into consideration) or not averaged (so that the speed of the user's speech should be taken into consideration). We will see which of these would be more beneficial.

Once we have found a way to compute a "useful" distance measure to judge a non-native speaker's pronunciation of a target phrase, we will endeavor to find the useful correlation between that objective measurement and the subjective notion of "good" or "understandable" pronunciation. We will determine this from the results of our study in which we have random participants of varying proficiencies pronounce phrases to be judged by native speakers. Incorporating this information into our pronunciation scorer, we should be able to offer the user feedback as to whether their pronunciation is improving, and with what proficiency "level" their pronunciation aligns. This may even be useful to extrapolate forward

and determine a rough estimation of the future performance of a given language-learner i.e., whether they will "be able" to learn a given language. Overall, the creation of such a tool, if truly language independent, will have great potential use in the domain of computer-assisted language learning technologies.

Milestones:

Milestone	*Date*
1. Build/Use an existing language independent recognizer to predict articulato features given acoustics.	10/11/13
2. Build/Use existing non-native recognizers for several languages to obtain phonem transcriptions in each language given acoustics.	10/11/13
3. Obtain recordings of native, nearly native, non-native and novice speakers for variety of different languages. Obtain these recordings for both male and femal speakers. (Ongoing throughout the recognizer creation process.)	11/8/13
4. Using Dynamic Time Warping, align the input speech with the target acoustic Compute an overall Mel Cepstral distortion value between the input speech and th corresponding target acoustics.	11/8/13
5. Using the phonemic transcriptions from step 2, compute average MCD value over each manner of articulation. Also, obtain an overall average MCD value t determine the overall pronunciation score. Furthermore, report manners c articulation for which the MCD value is high.	11/8/13
6. Using the DTW alignment from step 4, compare the durations of acoustics fo each phoneme in the input and target speech. If the duration of acoustics for a phoneme in the input speech is significantly higher than the duration of correspondi acoustics in the target speech, lower the pronunciation score accordingly and repo this issue to the user.	11/15/13
7. Perform the evaluation tests described in the 'evaluation' section below usin recordings from step 3.	11/29/13

Evaluation/Testing

The performance of the pronunciation scorer is evaluated using the following criteria:

1. Recognizing non-nativeness in speech

An ideal language-independent pronunciation scorer must be able to distinguish between native and non-native speakers by consistently assigning a higher score to native speech (independent of language). This can be tested by providing the pronunciation scorer with recordings of native, nearly native, non-native and novice speakers and recording the performance of the pronunciation scorer in each of these cases. The scores output by the pronunciation scorer are then compared to scores assigned to each

speaker by native speakers of a language.

2. Recognizing irregular length differences in pronunciation
If the phoneme duration for a speaker is significantly larger than the phoneme duration in the target speech then the pronunciation scorer must assign a lower score to the speaker and inform the speaker about this issue. At this point, no decision has been made about exactly how to test this. The project proposal will be updated once that decision is made.

3. Performance with male and female speech
The pronunciation scorer can be tested to see whether it handles male and female speech correctly by checking whether it awards the similar scores to male and female speech of similar competence. This can be done by providing the pronunciation scorer with recordings of native, nearly native, non-native and novice male and female speakers and checking the output of the scorer for male and female speech of similar competence.

Resources
Recordings (and corresponding transcriptions for each recording) from the following languages: Akuapem, Asante, Bambara, Bono, Dagaare, Dagbani, Dangme, Ewe, English, Fante, Ga, Gonja, Hausa, Jula, Kasem, Nafaanra, Nzema, Twi, Waale

Bibliography

Arun Kumar, Nitendra Rajput, Dipanjan Chakraborty, Sheetal K. Agarwal, Amit Anil Nanavati, "WWTW: The World Wide Telecom Web," NSDR 2007 (SIGCOMM workshop), Kyoto, Japan, 27 August, 2007.

Kominek, John, Tanja Schultz, and Alan W. Black. "Synthesizer Voice Quality of New Languages Calibrated With Mean Mel Cepstral Distortion." Language Technologies Institute, Carnegie Mellon University, USA, n.d. Web. 24 Sept. 2013. <http://www.cs.cmu.edu/~awb/papers/sltu2008/kominek_black.sltu_2008.pdf>.

Metze, Florian, Etienne Barnard, Marelie Davel, Charl V. Heerden, Xavier Anguera, Guillaume Gravier, and Nitendra Rajput. "The Spoken Web Search Task." N.p., n.d. Web. 24 Sept. 2013. <http://ceur-ws.org/Vol-927/mediaeval2012_submission_4.pdf>.

Metze, Florian, Xavier Anguera, Etienne Barnard, Marelie Davel, and Guillaume Gravier. "THE SPOKEN WEB SEARCH TASK AT MEDIAEVAL 2012." N.p., n.d. Web. 24 Sept. 2013.<http://www.cs.cmu.edu/~fmetze/interACT//Publications_files/publications/icassp2013-mediaeval-v2.pdf>.