Applause: A Learning Tool for Low-Resource Languages

Nikolas Wolfe, Vinay Vyas Vemuri, Lara J. Martin
Florian Metze
Alan W. Black
Language Technologies Institute
Carnegie Mellon University
{nwolfe, vvv, ljmartin, fmetze, awb}@cs.cmu.edu

Abstract
In this paper, we describe the concept of an interactive tool which can be employed to build a dialog system to facilitate pronunciation training in situations where only a few minutes of speech in the target language are available. We leverage recent advances in low-resource speech processing, and envision a tool which will help organizations working in the developing world to quickly create initial training lessons for new languages. Development will be possible using mobile devices alone, and without the requirement of significant technical skills or language-specific information. If users are ultimately able to acquire at least rudimentary proficiency in a low-resource language or dialect using our automatic system, they should find it much easier to communicate, establish trust, and build rapport with the local population than without such support. In this paper, we present the operational principles of our approach, describe a proof-of-concept implementation that we are currently developing, and summarize ideas for evaluation at the technical, language-learning, and user interface levels.

Author Keywords
natural language, low-resource, language learning

ACM Classification Keywords
H.5.2 [User Interfaces]: Natural language.

Introduction
Currently, interactive dialog and language learning systems are only available in a few, well-resourced languages for which speech data and language experts are readily available. By contrast, we want to investigate the feasibility of making a tool that can fundamentally perform the same functions in situations where only a limited set of recordings in the target language are available and with little to no additional human expertise.

The inspiration for this work stems from a project started by U.S. Peace Corps Volunteers in Ghana and Mali known as the Celebrate Language Audio Project, or CLAP [5]. ("Applause" is a play on this acronym.) The project was an effort to develop an efficient framework for the rapid development of audio lessons for low-resource languages. From 2010-2013, CLAP Volunteers collected an audio corpus comprised of 12 West African languages, (Akan, Bambara, Dagbani, Dioula, Ewe, Gonja, et al.) or about 36 hours of scripted, recorded speech.

In the following sections, we present a proof-of-concept pronunciation scoring application and its initial evaluation. In order to be applicable to as many different languages as possible, we used a language-independent approach to pronunciation scoring [4], similar to the approach proposed by Lee and Glass [1]. We focused on the extraction of audio features which are generally considered universal, such as articulatory features, fundamental frequencies, and a restricted set of Mel-Frequency Cepstral Coefficients (MFCCs). We also attempted to normalize in order to compensate for speaker-dependent features. A Classification and Regression Tree (CART) was constructed to predict pronunciation scores based on human-assessed training data. Our initial experiments were done using English data, verifying the feasibility of our fundamental approach before eventually pursuing the idea further with different languages.

Procedure

In our experiment, 30 utterances from 8 test speakers (a total of 240 utterances) were selected from the CMU Arctic Database [2], since accent labels and multiple speakers are available, making this data resemble our intended target audio. One additional speaker (rms, a "standard" American voice) was selected to represent the "target" pronunciation. The first three authors individually ranked all the utterances of the test speakers according to their similarity to the target speaker's pronunciation of the same utterance, disregarding gender and pitch. Starting with the first utterance and playing the speakers in a random order, the authors listened to the target utterance before hearing each source utterance. Utterances were allowed to be repeated, if necessary. These rankings were averaged together to produce a target pronunciation score.

In order to automatically compare the utterances, 26 articulatory features, 12 MFCCs, and the fundamental frequency (F0) were extracted from each recording, for a total of 39 features, using a 5ms frame step size. Higher-order MFCCs were discarded in order to minimize the influence of speaker-dependent features. Articulatory features were extracted using an Artifical Neural Network (ANN) [3]. Features included silence, voiced/unvoiced, vowel length features including diphthongs, long, short, and schwa, vowel height features including closed, mid, and open, and other vowel features including front, central, back, and rounding. For consonants, we extracted stops, fricatives, approximates, nasals, liquids, lateral approximates, labiodentals, bilabials, dentals, alveolars, post-alveolars, velars, and glottals. Each of these features was assigned a probability per frame.

Each utterance was then compared to its corresponding target pronunciation using Dynamic Time Warping (DTW) alignment based on mel-cepstral features, as is typically done in speech synthesis. The features were z-score normalized over all frames in the utterance in order to minimize speaker-dependent differences. A simple distance was calculated between the aligned target and reference frames, and finally the mean and variance were taken for each feature over the entire utterance, producing a single 78-dimensional feature vector per utterance. A CART tree was trained to predict the human-assessed pronunciation scores using these feature vectors.

Results

In order to evaluate the performance of the CART-based model, a 15-fold cross-validation was performed. For each fold, 2 utterances from 8 speakers (a total of 16 utterances) were set aside as test data, and the remaining 224 utterances were used to train 50 CARTs with stop

values 1 to 50. Among the constructed trees, the tree with the highest correlation was run on the test examples to generate predicted scores for that fold. Finally, the mean predicted scores for each speaker across all 15 folds were computed, and speakers were ranked in ascending order by their average scores. The predicted rankings were then compared to the human-assessed rankings, which were computed using the average human-assigned scores for each utterance. The results of this comparison are shown in Figure 1.

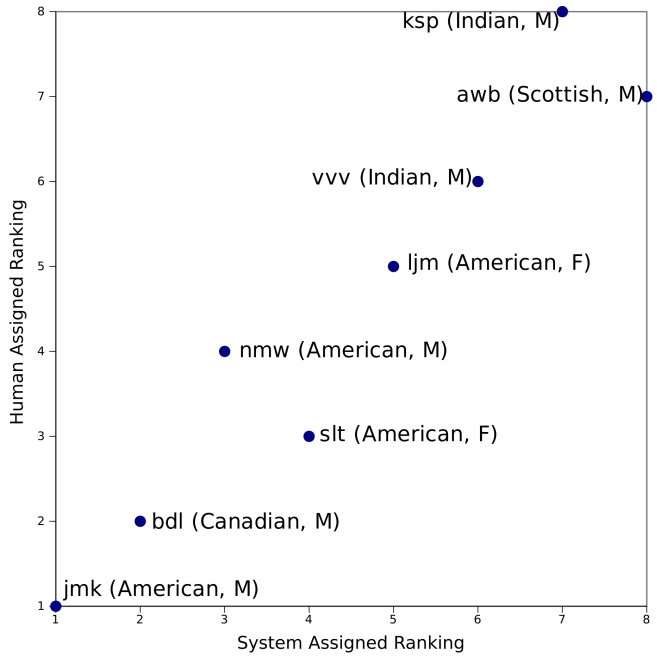

Figure 1: Averaged Human vs. Predicted Rankings with Accent Demographics. 2 speaker rankings are flipped in the automatic ranking. Overall "American-sounding" speakers receive lower ranks than other candidates.

In order to evaluate the applicability of our system in a dialog-based language training program, we conducted a phrase-matching experiment to validate whether, given a target utterance from a test speaker, the system can correctly identify the equivalent phrase from a set of utterances by a different speaker. For the 81 possible ordered comparisons between 9 speakers, we tested a single target utterance from the first speaker against a set of 15 (out of 30, to save computation) phrases from the second speaker. For 1,215 total comparisons, the system correctly assigned the target utterance a ranking equivalent to the minimum score 85% of the time (including ties). However, the system only assigned the overall minimal ranking to the target utterance 56% of the time. We think this is a useful baseline to improve upon in future work.

Discussion

Using only 240 utterances for training, we were able to construct a decision-tree based pronunciation scoring tool, which ranks speakers according to their perceived pronunciation "quality" with plausible accuracy. Fig. 1 shows that the manual and automatic ranking correspond well, and they intuitively "make sense," given the speakers' known accent. Of course, there is a chance our judging criteria are badly suited: should pronunciation quality be expressed as a rank within a group of eight speakers without allowing ties? For human assessment, it may have been a better to use a Likert scale. While a no-ties constraint ensures greater score variability, it does not account for the case in which all speakers perform equally, resulting in a skewed assessment. However, we feel that an absolute ranking is better suited to label low-resource languages without linguistic experts at hand.

One could argue that the scores generated by our system

are arbitrary, but an increased score over time may indicate an improvement in the user's pronunciation. Further research will be required to improve Applause and assess its impact on language learning and pronunciation. DTW can be used to perform speech recognition in small domains, so our scoring approach can directly support basic speech dialog. As discussed, we can already attest to the ability of our system to recognize utterances.

We consider the CHI Workshop an essential part of this effort, as a way to hear critiques, share ideas, and offer a novel perspective on the critical issue of creating useful tools for low-resource and endangered languages. Our research thus far has not focused on the important aspect of human-computer interaction (HCI) in automated pronunciation training, and we hope to gain some important insights in this regard from the CHI community.

Future Work

Since our research thus far has been limited to accented English, we first intend to experiment with data from actual low-resource languages, specifically our corpus of African language audio data. From CLAP resources collected by the first author, we have the foundations of a dialog system with which we plan to integrate our pronunciation scoring tool. We would like to work towards a distributable virtual machine that can be used to develop tools for any language.

Many questions remain: How can our system be made available to inexperienced users? Can lessons be designed by non-experts? Is language learning actually taking place here? To what degree is the learning process affected by differing user interfaces and feedback modalities? Is this training effective in the field? What would be the social impact of such tools to learn low-resource languages?

Summary and Conclusion

This paper describes a proof-of-concept implementation of the Applause low resource language learning tool. We described the fundamentals of operation, leveraging recent advances in low resource speech processing, and our initial tests using a well-understood data set. Given the amount and quality of data available, the system operates far from error-free, but our initial evaluation of technical performance characteristics are promising. We are looking forward to further exploring the challenges of this unique setting over the next couple of months, and to fully implement and evaluate the proposed system.

Acknowledgements

Special thanks to Prasanna Kumar Muthukumar for providing us with an implementation of the Articulatory Feature Extractor. We would also like to thank Maxine Eskenazi for her suggestions and advice.

References

[1] Ann Lee, J. G. Pronunciation assessment via a comparison-based system. *Proceedings of SLaTE 2013* (2013), 122–126.

[2] Black, A. W. Festvox: CMU_ARCTIC speech synthesis databases. `http://festvox.org/cmu_arctic/`.

[3] Black, A. W., Bunnell, H. T., et al. Articulatory features for expressive speech synthesis. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE (2012), 4005–4008.

[4] Precoda, K., Halverson, C. A., and Franco, H. Effects of speech recognition-based pronunciation feedback on second-language pronunciation ability. In *Proceedings of InSTILL 2000*, University of Abertay, Dundee, Scotland (August 2000), 102–105.

[5] Wolfe, N., et al. CLAP: The Celebrate Language Audio Project. `http://celebrate-language.com/`.