Audio-Based Practical Language Learning: A Platform for Crowdsourced Data Acquisition and Personalized Education

Nikolas Wolfe, Bhiksha Raj

Language Technologies Institute, Carnegie Mellon University

{nwolfe, bhiksha}@cs.cmu.edu

Abstract

Language learning applications are among the most popular ways in which people are employing smart devices for self-directed education. While well-established services such as Duolingo, Pimsleur, Rosetta Stone, Babbel, Busuu et al. compete for market share and teach according to varying theories of L2 acquisition, few if any of these applications can claim to offer automatically generated pronunciation and intelligibility feedback in a way that effectively prompts and assists a learner to improve their oral abilities in a given language. It is generally agreed that automatically generating effective pronunciation feedback is a hard problem requiring sophisticated machine learning and signal processing techniques as well as large amounts of annotated data. While recent advances in low-resource speech technology are noteworthy, we argue this is short-sighted and that the prevailing experience in modern machine learning is still that there is no data like more data. We therefore propose a divergent approach in which we emphasize speaking skills and immersion-style L2 acquisition, and argue that a concerted crowdsourcing campaign to gather the data required to do this in increasingly automated ways is an integral long-term component of any successful language learning platform or application.

1. Introduction

Effective tools to assist second language (L2) aquisition have been the subject of much research due to their wide practical applicability [**?**] [**?**] [**?**]. From easing cross-cultural and interpersonal communication in travel, business, international relations and other professional domains, the development of a truly automated and scientifically robust language learning platform has become something of a holy grail in the field of language technology. Part of the mystique of this problem stems from the fact that we still have only rudimentary models for the actual neurological process of human language learning and understanding, and as such there are many competing theories and heuristics for practical L2 acquisition [**?**] [**?**].

The preeminent approach taken by popular computer-assisted language learning (CALL) applications such as Rosetta Stone [**?**] and Duolingo [**?**] [**?**] favors a didactic, linear pedagogy in which language is learned as an abstraction where simple grammatical rules and core vocabulary are iteratively combined and expanded upon to teach more complex linguistic concepts over time. Speech training and assessment are something of an afterthought for these systems, the implied assumption of course being that literacy in a given language is more important than oral competence. From a practical perspective however, it should be axiomatic that the most readily applicable and expedient competencies in L2 acquisition are not reading and writing, but rather listening and speaking skills [**?**] [**?**] [**?**] [**?**] [**?**].

In contrast to this is the more orally immersive, memory-reflex oriented structure of a Pimsleur method guide [**?**] [**?**] [**?**] [**?**] or the collaborative, community-oriented approach of Busuu [**?**] [**?**] or the promising but ill-fated LiveMocha [**?**] [**?**]. Applications such as Carnegie Speech's NativeAccent® [**?**] and Babbel employ more advanced speech processing techniques to assess pronunciation "nativeness" and provide corrective feedback. A large amount of academic research has also been done in the field of computer-assisted pronunciation training (CAPT) and the potential use of speech recognition for language learning [**?**] [**?**] [**?**] [**?**] [**?**] [**?**] [**?**] [**?**] including notable systems for teaching English, French, Mandarin [**?**] [**?**], Hindi [**?**], and Arabic [**?**] among others [**?**] [**?**].

While an exhaustive evaluation of the pros and cons of each of these systems is beyond the scope of this paper, we can broadly categorize their approaches as being focused on the detection and classification of pronunciation errors with the contextual aim of helping a learner develop a "native-sounding" accent. Learner feedback in systems like NativeAccent® is geared towards emphasizing native articulatory correctness, whereas most systems simply attempt to classify common mispronunciation errors and offer numeric scale assessments, e.g. a score between 0 and 100. It is arguable that offering numeric feedback is the effective equivalent of asking a learner to self-assess, which is an error-prone process at best [**?**].

A poignant, though perhaps simplistic critique of typical approaches to CAPT is that so-called "nativeness" is both a culturally loaded and ill-defined concept. Many of the people in the world who communicate in English do not speak with a British or American accent, and furthermore, why should they? It is arguable that intelligibility and understandability are better and more useful benchmarks to define. There are of course situations in academic or professional environments where grammatical and articulatory correctness are perhaps important, but this is a high bar to set for most practical purposes. In day-to-day communication or travel situations where the only metrics of success are showing respect and being properly understood, it is typically the effort that counts more than the place of articulation. We therefore suggest that a learner's oral intelligibility by first-language (L1) speakers be the gold standard by which any practically-oriented CALL or CAPT system is evaluated.

2. Discussion & Methodology

In order to design and construct a language learning application which is oriented towards practical speaking skills, we must evaluate competing methodologies for L2 acquisition and pragmatically choose the most well-suited attributes of each. Trivially, we have chosen to favor a CALL/CAPT application over the design of a new classroom-based curriculum or other tra-

ditional approach. Computers in general are uniquely suited for language training because they have infinite amounts of patience and well-established memory-enhancing methods such as spaced repetition [**?**] [**?**] [**?**] [**?**] are relatively easy to implement programatically.

Unfortunately, computer systems still lack the eponymous ability of trained human langauge instructors to make pointed assessments of speech intelligibility. Machine-based methods in this domain still critically lag human performance and as such we argue humans cannot be cut out of the loop yet. Part of this is undoubtedly due to a lack of large annotated pronunciation assessment corpora for supervised learning. Automatic speech recognition (ASR) technology has been around for decades, and yet only recently with the infusion of massive amounts of annotated data and proportionally massive supercomputing systems has ASR performance in languages like English and Mandarin begun to equal or outstrip human performance [**?**] [**?**] [**?**].

Pronunciation and intelligibility assessment can be mechanistically described as process of assessing an input speech signal and generating human language to describe how to improve or change the signal to better match some target speech signal(s). To offer a practical example from a Pimsleur-style language guide for Akan:

Narrator	Listen to the phrase for "I'm fine" or "I feel fine"
Male Speaker	Me ho y

In the domain of generating human language to both classifiy and describe signal inputs like images, recent research argues that this is not only possible, but likely already within reach [**?**] [**?**]. Automatic methods for image captioning, unsurprisingly, rely on large amounts of input images annotated with descriptions and typically a large and deep neural network to figure out the relationship between them. It is hardly reaching to suggest an analogous process for automatic speech pronunciation and intelligibility assessment.

Moreover, in the domain of human language learning, it is often colloquially asserted that immersion training is the best way to practically learn a language. This is approach is soundly rooted in both theory and practice and is used in language training by organizations such as the U.S. Peace Corps and Foreign Service [**?**] [**?**] [**?**] [**?**] [**?**]. The practical academic discipline of field linguistics has also established a rigorous and broadly applicable methodology for discovering rules and structure in unknown languages [**?**] [**?**] [**?**].

There are certainly aspects of popular language learning platforms, immersion training approaches, field linguistics, and modern language technologies which can be constructively combined to make more effective oral language learning software. There is also a wealth of existing language resources from popular platforms which can be bootstrapped once a general framework is established. A resource which has only recently started to receive much attention for filling in the gaps, however, is crowdsourcing [**?**] [**?**] [**?**].

Using the wisdom of the crowd [**?**], many problems which today seem intractable can be approached incrementally. Automatically assessing pronunciation and intelligibility and generating feedback that assists people improve their oral skills in a given language in an empirically verifiable manner is well known to be a hard problem for computers which is relatively easy for human beings. It follows that crowdsourcing the work of pronunciation assessment is a good way to gather annotated data which can be used to iteratively improve machine models over time. We may not know the best automatic methods for generating feedback today, but it is certainly a worthwhile investment of time to begin gathering data for the future. Much like earning incremental interest and returns on an initial monetary investment, the passive gathering of data over time will reap dividends down the road.

3. Proposed System

With the previous discussion in mind, the general language-learning platform we propose has three main sub-systems:

1. An intelligent spaced repetition based mobile application focusing on oral recall, bootstrapped with data from existing language learning courses and applications
2. A crowd-based collaborative data acquisition service to assess pronunciation and intelligibility of learners' speech in order to gather constructive feedback, and,
3. A systematic means by which a learner can be guided to figure out a language for which only limited or insufficient learning materials are currently available

We can consider each of these items in more detail in turn.

3.1. Mobile Education App

The rise of mobile as a platform for self-directed education dictates that any new language learning application be built for smart devices from the get-go. At the same time, given the known efficacy of immersion training for oral language learning, we must consider how to simulate this sort of teaching style on a smart device. Immersion training, by definition, is a teaching method in which the target language is the medium of instruction [**?**]. While the strictness of this rule is debatable, the general idea is that learners are prompted to listen, comprehend, and speak in the target language from the very beginning. From an application development perspective, audio must therefore be the primary medium of communication.

Pimsleur guides are better poised than didactic regiments like Duolingo for immersion training because they are oriented towards practical scenarios and are almost entirely in audio format. (In fact, Paul Pimsleur developed one of his original compact language guides in Akan for the Peace Corps in 1971, and the first author both used this guide as a Peace Corps Volunteer in Ghana and extended its basic structure to build additional language guides for 13 West African languages [**?**] [**?**] [**?**]). Of course, the lack of intelligent corrective feedback in Pimsleur guides is an admittedly critical flaw.

Both Duolingo and Pimsleur employ spaced repetition, though clearly intelligent repetition based on learner performance and confidence is preferable to a simple audio recording. Duolingo, notably, also employs the crowd for translation tasks and attempts to make the interaction fun [**?**] [**?**]. Independent evaluations indicate that Duolingo's enjoyability is crucial to curbing learner attrition rates.

With a focus on immersion-style training using audio and borrowing useful attributes from existing language learning platforms, it hardly makes sense to try and reinvent the proverbial wheel with respect to content for most major languages. Language learning guides for major European and Asian languages (the so-called "big" languages) are practically ubiquitous. Of course, while content may be plentiful, variations in pedagogy are more scarce. Thus, the first step is to gather existing data and restructure it into a better pedagogical format.

3.1.1. Bootstrapping Existing Language Resources

One crucial advantage of a Pimsleur-style guide for an immersive language training experience is the choice of content itself. Pimselur lesson content is chosen specifically to match real-world scenarios and order is based more on practical utility than grammatical complexity, e.g. phrases like "Good morning" and "I don't understand" are taught before simple nouns like "boy" and "cat." Determining the correct ordering of practical introductory phrases will likely change for different languages, however it is unlikely that the words "boy" or "cat" will ever occur before simple greetings in a real-world setting. The structure of a Pimsleur guide can thus be effectively bootstrapped to create a modifiable template for most introductory language lessons, as implemented by Wolfe et al. for the Celebrate Language Audio Project (CLAP) [**?**].

Additional resources to bootstrap into a practical language learning application include the guides created by the U.S. Foreign Service Institute, the Defense Language Institute and the Peace Corps, all of which are in the public domain and freely available online [**?**] [**?**] [**?**]. University resources such as the Indiana University Center for Language Technology (CeLT) and the UCLA Language Materials Project et al. are also freely available online resources for language technologists to use [**?**] [**?**]. Using only the resources listed above, we were able to amass nearly 700 gigabytes of data for 157 languages, totalling 29,227 raw audio files.

3.1.2. Building Structured Lessons From Disorganzied Data

3.2. Crowd Data Acquisition and Speech Assessment

3.3. Guided Language Discovery