James Salsman
Senior Engineer, EF Learning Labs, Shanghai, China
james.salsman@ef.com or jps@talknicer.com

Development challenges in automatic speech recognition for computer assisted pronunciation teaching and language learning

Biodata

James Salsman is a Carnegie Mellon-trained statistician working on pronunciation assessment using automatic speech recognition for language learning since joining Stanford's Applied Speech Technology Laboratory as one of its first employees in 1996. His interests include the use of diphone scoring, correct scoring by adaptation to accents and dialect, automating the production of computer aided pronunciation teaching software with crowdsourcing, and open source solutions to encourage development of new and innovative self-study solutions. He has been working at Education First’s EF Learning Labs in Shanghai since early 2013.

Abstract

Automation has been improving the productivity of almost all human endeavors, recently including language instruction, but also has been displacing human labor, and threatens to exacerbate limited but pedagogically essential student-teacher interaction opportunities and teacher demand. Several methodological challenges in the development of automatic speech recognition for computer assisted pronunciation teaching (CAPT) range from the need to support instead of displace human teachers, software engineering issues, the availability of appropriate technology to support clinical speech language pathology applications, addressing vastly larger numbers of less literate students, and the use of open source self-study applications to attract and retain students for human tutoring. Each of these challenges can be addressed, but potential solutions range widely in difficulty, time and resource requirements, costs and benefits. Software engineering issues such as testing and validation of enhancements and new features have occupied more time and required more effort than most of our other development issues combined. The use of physiologically neighboring phonemes, diphones, and segment durations are all likely to help improve learner outcomes. The language instructor’s experience of computer-assisted pronunciation assessment can be enhanced by offering comparisons of students’ utterances to exemplar pronunciations for each of those attributes.

Short paper

Technology improvements are often a mixed blessing. While nearly every aspect of our working lives has been assisted by automation, such advances have not always resulted in increased standards of living. Technology improvements are contractionary in the economic sense (Basu, Fernald, and Kimball, 2006.) In much of the developed world, productivity gains have not been reflected in real wage and purchasing power increases
since 1973 (Mishel, 2012.) Computer aided instruction has often allowed increased class sizes effectively similar to earlier smaller classes, along with greater opportunities for distance education. But many education leaders wisely keep class sizes small to remain competitive in the rising tide of open educational resources, further technological developments, and competitive pressures of globalization (Blatchford, Bassett, and Brown, 2011; Konstantopoulos and Chung, 2009.)

More steeply progressive taxation and workweek length reduction can offset contractionary economic pressures (Griffith, 2004; Frank, 2003; Frank, 2005; Pickett and Wilkinson, 2011; Finkel, 2010; Golden and Glosser, 2012.) However, because of the lack of progress in these areas — possibly due in part to the failure of educators to communicate salient facts sufficiently well (Kornhauser, 2004) — student-teacher interaction opportunities and teacher demand are both at risk from institutional downsizing due to government austerity, (Ranieri and Almeida Ramos, 2013; Anand, Mishra, and Peiris, 2013) curriculum standardization, educational institution and publisher consolidation, and the rush to massive automation without careful evaluation of long term outcomes (Oremus, 2013.)

Interactions between students and teachers are particularly important for spoken language learning, which involves microsecond scale vocal tract motor coordination and dexterity, the interaction of long-term and short-term memory, modeling the understanding and expectations of conversants, real-time parsing and disambiguation, with simultaneous listening, patience, and related linguistic, intellectual and emotional skills. Immersion is widely recognized as far more effective for spoken language learning than self-study or large classes, even with the latest technology. For example, Carnegie Speech Company’s NativeAccent computer assisted pronunciation teaching (CAPT) system may provide diminishing returns relative to small class instruction after about 40 hours of use (Carnegie Speech, 2013.)

**Current practice**

At EF Learning Labs, we offer automatic speech recognition technology for self-study spoken English language speech and pronunciation instruction to hundreds of thousands of adults and children on a variety of platforms including personal computer web browsers and mobile devices (EF Learning Labs, 2014.) We plan to offer spoken language instruction in other major Indo-European and tonal languages on at least the same wide variety of platforms.

EF also offers group language lessons and private one-on-one spoken language instruction over teleconferencing and in person in hundreds of offices around the globe. So our corporate leadership is well aware of the importance of small class size, and those of us working in computer aided instruction at EF greatly benefit from the expectation that we strive to support instead of displace human-led spoken language instruction.

**Development challenges**

Software engineering challenges that we have encountered in development involve cross-platform support, resource allocation for building and testing new features, and modifying and testing existing features for enhancement. Software patents often sharply limit development options increasingly throughout the world (Lee, 2013.)

Because pronunciation quality is partly aesthetically subjective and listener intelligibility is subjective, dependent on context and background noise, and a matter of degree, panels of expert judges often disagree on pronunciation quality and intelligibility questions at the phoneme, word, and utterance levels. So we also face design and accuracy challenges regarding measurement of speech segment confidence scores as well as durations, score aggregation, motivational tactics, (del Soldato and du Boulay,
1995) and validation for a wide variety of learner backgrounds, situations, ages, and skill levels. While our primary goal is education first and foremost, the production of excellent CAPT educational services demands accuracy and robustness suitable for use in clinical speech language pathology applications (e.g., Middag, Martens, Van Nuffelen, and De Bodt, 2009; Baghai-Ravary and Beet, 2013.)

Addressing vastly larger numbers of less literate students in developing regions involves the use of limited memory bandwidth and very efficient automatic speech recognition on earlier generations of mobile devices (Sensory, Inc., 2013.) We believe that providing spoken language educational opportunities to the greatest number of learners without displacing human instruction can be accomplished with branded open source self-study applications to attract and retain students for human tutoring (Ronanki, Li, and Salsman, 2012.) But open source automatic speech recognition systems are sometimes less efficient than commercial alternatives, especially on earlier generations of mobile devices.

**Discussion**

Each of the challenges identified above can be addressed, but potential solutions range widely in difficulty, time and resource requirements, costs and benefits. Over the past several months of 2013 and 2014, testing and validation of new features and related software engineering issues have occupied more time and required more effort than most of our other development issues combined.

To best support language instruction, we have been developing the use of physiologically neighboring phonemes, i.e., sounds produced with similar vocal tract articulations, to identify and discern between serious mispronunciations and incidental errors (Ronanki et al., 2012.) We are using diphones, i.e. the last half of one phoneme followed by the first half of the next, as alternatives and supplements to phonemes and triphones for both automatic speech recognition and pronunciation scoring (Cole, Serridge, Hosom, Cronk, and Kaiser, 1999.) We plan to model learner fluency and select the sequence of self-study practice exercises using cumulative diphone scores. We are scoring segment durations to indicate syllables and words pronounced too quickly relative to exemplary pronunciations. We have measured substantial potential improvements from all of these techniques. We hope to integrate all of them with our products and services in 2014.

The language instructor’s experience of computer-assisted pronunciation assessment can be enhanced by offering comparisons of students’ utterances to exemplary pronunciations in ways that illustrate the measurements of physiologically neighboring phonemes, diphones, and speech segment durations. For example, mispronunciations might be annotated with International Phonetic Alphabet symbols for both the expected pronunciation and its physiologically neighboring phoneme which most closely matched the observed speech. Diphones can be used to highlight difficult phonetic transitions, for example when two adjacent phonemes are both mispronounced. Duration scoring can annotate not just words and sub-word segments given insufficient emphasis, e.g. such as might confuse “fourteen” with “forty,” and can highlight missing glottal stops essential to discern, for example, “harder” from “hard or.”

Because the learner shares the instructor’s role during self-study, we do not anticipate substantial difficulties in offering these enhancements to instructors as well as independent learners. But the reality of software development demands lengthy cycles of engineering, debugging, quality assurance testing, usability testing, release cycles, documentation, and instructor training.

Such mundane software production challenges are faced by most if not all CALL developers, and appear to present the largest problems to those of us trying to support educators with emerging automatic speech recognition technologies for pronunciation.
assessment. So we strive to improve our software engineering processes with new techniques, recently for example by exploring the integration of our Agile software development methodology with the Win-Win requirements negotiation model (Khan, Wahab, and Saeed, 2014.)

References


