International Conference
ICT FOR LANGUAGE LEARNING
7th Edition

"Mira, Mamá! Sin Manos!" Can Speech Recognition Tools Be Soundly Applied for L2 Speaking Practice?

Thomas Plagwitz
Charlotte, NC / USA
plagwitz@outlook.com

Abstract

Based on a recent Language Resource Center (LRC) and departmental implementation (see e.g. Workshop), this paper gives a brief overview over the current ***research status*** *of automatic speech recognition (ASR) - i.e. "to convert speech into a sequence of words by a computer program" - and comparison of current product* ***implementations*** *more or less readily available to the end user, discussing their possible practical* ***applications in second language acquisition (SLA)*** *programs for speaking practice. We demonstrate how currently widely available ASR technology for 5 to 7 (depending on definition of language variants) of the most popular languages in current SLA can be implemented in many SLA programs around the world in an economically (practically no extra cost other than for local installation; individual, short, flexibly scheduled speech assignments mean no need for funneling entire classes hogging the LRC) and pedagogically viable way. We used Windows-7 Enterprise (and up) ASR (likely available to you already in your LRC) with Microsoft Language Packs (free) which give you high-quality ASR for a subset of display languages. However, this still required not only carefully* ***controlling the technical aspects*** *that make such projects often fail in language programs with their limited resources, but also* ***managing expectations****, not to replace the teacher, only blend AI with human intelligence in order to widen the "expert bottleneck" for the learner and relieve the teacher from routine tasks, like scheduling meetings (even if online) and/or listen to the entirety of a student's oral production, while allowing for human personalized feedback where needed. We also carefully* ***designed the task****: Teacher creates speaking practice assignments based on syllabus, correlated with assigned textbook activities and with corresponding upload assignments in the (optional) LMS (email could replace). Easy integration of speaking practice tasks into the syllabus is a major advantage over the usually intractable problems with progression when trying to integrate language learning material providers' ASR solutions if the course syllabus is not primarily based on those materials. For the speaking* ***practice tasks****, we link both training and documentation for end users and* ***sample ASR task completions*** *(as* ***screencasts****) by teachers and learners, to demonstrate actual recognition quality (and alert to minor pitfalls). We discuss* ***assignment variations*** *(immediate student reaction to bad recognition by using of built-in speech correction; simpler GUI control through speech, based on a more restrictive vocabulary) and finally recommend integration of the task results into* ***ePortfolios*** *to show off language learner achievement.*

1. Introduction

ASR approaches the "plateau of productivity" in Gartner's current hype cycle [1] – time to revisit the erstwhile "Holy Grail" of CALL, deemed infinitely promising [2], a computer capable of 'understanding' human language? Disappointed by remaining "silly" recognition errors, many, including SLA programs, banned ASR (also my own former Auralog Tell-me-more (ATMM) installation) into the "valley of disillusionment". Yet despite "[s]peech [being] a highly variable signal" ([3], 75, [4], 356f.), ASR, since introduction of the "statistical paradigm" in the 70s still dominant today ([4], 339, 351), has seen rapid progress, owing to improvements in infrastructure (Moore's Law, larger speech corpora for modelling), knowledge representation (searchable unified graph representations that allow multiple sources), language models (based on n-grams) and algorithms (HMM, ML during system training), search (stack and Viterbi) and automatic metadata tagging for topic and speaker [3]. Government sifting through massive "open" online data, including speech, and ubiquitous hyper-personalized mobile devices that rely more on – and produce more - speech for our corporate "overlords" make further progress likely ([4], 358f.). Consequently, no sooner than pronounced a dead horse [5], ASR rebounded with major breakthroughs both in R [6]&D [7]. What cheap ASR benefits can an LRC reap now, and with which pedagogy, hopefully to brace the upcoming ASR deluge?

2. SLA needs and LRC environment

My use of ASR integrates with **LRC activities** for oral practice, assessment and ePortfolio in a digital audio lab 359yUV (expand codes to e.g. http://goo.gl/359yUV). The vital need for ASR was however

our **1st-year Spanish program** conversion to “hybrid” mode: 50% contact hours, more online textbook exercises, but only self-grading, not the integrated speech recorder. "Paying special attention to the development of oral proficiency" is vital in hybrid classses to "help compensate for the reduced face-to-face classroom interaction" [8]. The obvious standout feature of ASR - not only for Hybrid Spanish - is **immediate automated feedback** via transcription “response”. Crucial for Spanish was also the **wide availability** of **Windows-7 Automatic Speech Recognition** (W7ASR) speech practice activities in our LRC (like many, limited in seats, licenses, hardware). W7ASR’s only requirements are Windows-7 and headphones, of which we had twice as many as Sanako, 6 as many as webcams. Worse: Our 1st-year course caps (initially even the smaller 2nd-year) exceed our Sanako licenses, and even 2nd-year Spanish sections (half of 1st-year!) LRC capacity to schedule a final oral exam. Only a more flexibly scheduled W7ASR homework assignment can reach all students with guided speaking practice.

3. ASR Tools

Gartner undoubtedly considered Siri, Google Now and Cortana - requiring much more AI than ASR only -, but mobile phones won't get main-streamed here soon, have power and bandwidth issues for continuous ASR, more limited use cases (search) than suitable for SLA and likely (or else we may have to worry about FERPA) breadth- and quality-constraints of "speaker-independent ASR system[s]" ([9], 255). For many institutions, most **alternatives** are precluded by cost. The free browser-based Google-Translate voice-input is, as of this writing, only available in English (and many of our teachers dread it, for students routinely abusing it as machine translator, not dictionary). Instead, we installed all available language technology add-ons in Windows-7 (and Office 2010: overview cya3eG; all easier in Windows-8 [10]), free for Windows Enterprise with volume licensing (most common at educational institutions). A few of these – for "it takes a global village" of local PC installations to warrant the cost -, but still covering over 6/7th of LRC enrolment, namely the Chinese, English, French, German, Japanese and Spanish **language packs** for the Windows-7 MUI, include **W7ASR** which, though underrated and -advertised, consistently ranks next to market leading and self-advertising Dragon Naturally-Speaking [11] [12].

4. Tool limitations and affordances

W7ASR was **not designed for SLA**, **accents** pose a big problem for ASR [13], and despite work (see e.g. [14]), “except for small gains, the problem is largely unsolved” ([4], 357) – which, however, increases the incentive for “good” pronunciation. Still, many promising studies on ASR for SLA ([15]), [13]) used Dragon - not an SLA software either. I found W7ASR sufficient for what I tried, unlike the SLA-specific ATMM ssctfc. W7ASR does not present the user with an **aural model** to match phonetically, nor **voice graphs** (could be implemented). However, visual feedback for pronunciation improvement is contentious ([16], 317ff), with only experimental, limited (prosody ([17], 464), phoneme/word-level or pitch) gains, based on heavily simplified "visual feedback systems" ([18], 40), and intense "additional support from the instructor" ([19], [18], 134) – exactly what we need to avoid. After observing students using ATMM's voice graph, I concur: "Spectrograms are uninterpretable to non-experts" ([20], 747). W7ASR also lacks **learning content** – a blessing in disguise: Despite considerable systems integration efforts, ATMM's ASR was barely used since ATMM's content could not be aligned with existing syllabi and textbooks. ATMM seemed largely a "**speaker-independent"** ASR system ([9], 255), despite individual accounts without obvious adaptive voice-training - except for ESL learners with especially poor (to me incomprehensible) pronunciation, after several non-improving attempts, suddenly being "waved through". For good recognition results, "software has to be trained to each individual speaker’s voice" ([15], 20). Siri, Google Now or Cortana may hide it. But **speaker-dependent** W7ASR sends students through a voice-training before the first assignment (and later also learns behind the scene).

4. Task implementation (prerequisites installation, configuration, training)

Basically, students record screencasts of dictation into MS-Word and correct recognition, tracking changes (training and documentation for your LRC and ASR samples: hLFGQt (pwd:rwsa7)).

5. Task design

W7ASR has 2 modes: **Voice command** (VC) and Dictation. The former operates the GUI with voice replacing mouse - like ATMM's "**[c]losed-response** design [,] display[ing] a few utterance choices for learners to say" ([21], 286 - rather restricting ([17], 464) and not improved by poor accuracy Lio9tm, nor by their new owner's algorithm ([20], 747) -, albeit with more choices (and user-extensible, with

Windows speech macros), most not displayed (see cheat sheet fzlVfQ). **Dictation** mode replaces keyboard with voice and is a large-vocabulary or "**open-response** design" ([21], 287), "response" meaning: typing out speech input, as a "**continuous** ASR system" ([9], 250). Microsoft advised to speak naturally and fluently and not over-enunciate - I upped it a bit. W7ASR can return to "isolated word" ([9], 253) or "**discrete** word ASR" ([9], 255) - not trivial, given the inner probabilistic workings, but useful during corrective passes over misrecognized words (NfvwKO, @"User corrects with speech").

Crucial TCO-wise is the easy integration of W7ASR with the **textbook-based** syllabus (until online textbooks add - like audio players, then recorders - ASR). Given that reading aloud is being reappraised as an L2 learning strategy [22], my **sample tasks** have advanced students read out authentic reading-for-the-gist textbook assignments, or beginners practice discrete speaking instead of writing within cloze-exercises, flip the classroom to homework speaking practice, from beginner's conversation phrase templates to the Miranda-warning (convey to ASR → suspect) of our "Spanish for Law Enforcement", and the most advanced speak instead of write essays (details and samples: mgpJN7).

W7ASR can not only grow on you, but with you, and accompany learners throughout their studies. Students are encouraged to keep updating their **speech profile**, for W7ASR continuously learns, and collecting task screencasts for their **ePortfolio** – show employers, unable to judge your accuracy, dictating a letter in L2 without proofing errors. Our task design is **multifaceted, multimodal and reinforcing**, combining all 4 skills: listening (to a model or the own pronunciation), (re-)reading (source text and recognized text), speaking, and some corrective writing. Not quite gamification - but having a 1st-year/-time language learner talk, in privacy, and control a machine with “incantations” makes W7ASR look like the once-envisioned tool that **encourages "[c]omprehensible output** (…) - [which] must be produced with the expectation that it is going to be 'understood.' Under these conditions, the learner is expected to attempt to use target language forms that may stretch his or her competence." ([23], 27)

6. Grading & Conclusion

"Speech recognition is not about building HAL9000. (…) Our job is trying **to find a good use** of an imperfect, often crummy, tool that can sometimes make our life easier." [24] W7AST – while anything but crummy, cf. my linked SLA samples - can neither “understand” nor grade students. It won’t replace, but relieve the teacher - a different twist on "it is uncertain that the same improvements [achieved using ASR] could not have been achieved through traditional classroom-based pronunciation instruction" ([20], 748). For **grading support**, we provide the teacher both maximum simplicity and, if desired, additional evidence: All speaking assignments are documented through uploaded screencasts of speech, its transcription and manual correction. Grade for the practice effort on submission of a screencast alone; or on the number of corrections, visible, thanks to “**track changes”,** in the last frame. Or, if questioning ASR validity, rewind to misrecognized words. Or, for helping with LMS assignment comments, recheck the entire screencast – student will thank you for additional, after immediate feedback. However, no need to closely assess actual pronunciation accuracy, if we accept W7ASR’s language model as a cost-free approximation of what constitutes "comprehensible output". It has long been established that "speech-interactive CALL [can] use speech recognition for building confidence and fluency rather than for pronunciation assessment". "The simplest [learning] principle is **practice makes perfect**. (…) A machine that listens is a good practice tool” ([9], 290). W7ASR could be that machine, available for free in your LRC.

References

[1] Babbage, Charles. Difference engine. Divining reality from the hype. The Economist. 08 27, 2014.
[2] Tutors That Listen. Holland, V. Melissa. 1999, Calico , Vol. 16, pp. 245-250.
[3] Research Developments and Directions in Speech Recognition and Understanding, Part 1. Baker, Janet and others. 2009, IEEE Signal Processing Magazine, pp. 75-80.
[4] Huang, Xuedong and Deng, Li. An Overview of Modern Speech Recognition. [ed.] Nitin Indurkhya and Fred J. Damerau. Handbook of Natural Language Processing. 2. 2010, pp. 339-366.
[5] Fortner, Robert. Rest in Peas. Posterous. [Online] 2010. http://robertfortner.posterous.com/the-unrecognized-death-of-speech-recognition.
[6] Seide, Frank and others. Conversational Speech Transcription Using Context-Dependent Deep Neural Networks. INTERSPEECH. 2011, pp. 437-440.
[7] McMillan, Robert. How Google Retooled Android With Help From Your Brain. Wired. [Online] 02 18, 2013. http://www.wired.com/2013/02/android-neural-network/.

[8] ADFL. Suggested Best Practices and Resources for the Implementation of Hybrid and Online Language Courses. [Online] ADFL , March 2014. http://www.adfl.org/resources/resources_Hybrid%20and%20Online%20Language%20Courses.htm .
[9] Software That Listens. Wachowicz, Krystyna and Scott, Brian. 1999, Calico, Vol. 16, pp. 253-256.
[10] Hamilton, Ian and Sinofsky, Steven. Using the language you want. Building Windows 8. [Online] Microsoft. http://blogs.msdn.com/b/b8/archive/2012/02/21/using-the-language-you-want.aspx.
[11] Trigon. Windows 7 Features: Speech Recognition VS. Dragon Naturally Speaking. [Online] 4 13, 2010. http://trigon.com/tech-blog/bid/32089/Windows-7-Features-Speech-Recognition-VS-Dragon-Naturally-Speaking.
[12] Anderson, Nate. Win 7's built-in speech recognition: a review. Arstechnica. [Online] June 1, 2010. http://arstechnica.com/information-technology/2010/05/win-7s-built-in-speech-recognition-a-review/.
[13] Interlanguage speech recognition by computer. Mascia, Rita and Selinker, Larry. 2001, Apples, Vol. 1, pp. 19-55.
[14] Construction of a Rated Speech Corpus of L2 Learners' Spontaneous Speech. Yoon, Su-Youn and others. CALICO, Vol. 26, pp. 662-673.
[15] The Use of Speech Recognition Software. Coniam, David. 1998, Calico, pp. 7-23.
[16] Learning French Pronunciation. Knoerr, Alysse and Weinberg, Hélène. 2003, Calico, Vol. 20, pp. 315-336.
[17] Using a Computer in Foreign Language Pronunciation Training. Eskenazi, Maxine. 1999, Calico, Vol. 16, pp. 447-469.
[18] Teaching Tone and Intonation With Microcomputers. Chun, Dorothy. 1989, Calico, Vol. 7, pp. 21-46.
[19] Eyespeak [Review]. Tao, Rui. 2007, Calico, Vol. 25, pp. 126-136.
[20] Computer Assisted Pronunciation Training. Thomson, Ron. 2011, Calico, Vol. 28, pp. 744-765.
[21] Speaking. Egan, Kathleen B. 1999, Calico, Vol. 16, pp. 277-293.
[22] Reading aloud. Gibson, Sally. 2008, ELT Journal, Vol. 62, pp. 29-36.
[23] Multimedia CALL. Chapelle, Carol A. 1998, Language Learning & Technology, Vol. 2, pp. 22-34.
[24] Pieraccini, Roberto. Un-rest in Peas. Blogspot. [Online] 05 2010. http://robertopieraccini.blogspot.com/2010/05/un-rest-in-peas-unrecognized-life-of.html.
[25] Spehr, Michael and Wiseman, Raymond. Betriebssystem: Auch Vista bleibt Windows. FAZ. [Online] 01 23, 2007. http://www.faz.net/aktuell/technik-motor/computer-internet/betriebssystem-auch-vista-bleibt-windows-1411370.html.
[26] Entre dicho y hecho …. Lafford, Barbara A. and others. 2007, Calico, Vol. 24, pp. 497-529.