Deep Language Learning

Steven Greenberg

Address:
Silicon Speech
17270 Greenridge Road
Hidden Valley Lake,
California 95467
USA

Email:
steven@siliconspeech.com

How technology enhances language instruction.

Technology Is Transforming How Students Learn a Foreign Language

In *Star Trek*™ and other science fiction, alien civilizations communicate via a universal translator that seamlessly translates from one tongue to another, making alien-language instruction superfluous. Unfortunately, such flawless universal translation is unlikely to arrive anytime soon (despite recent advances).

So, at least for now, the most effective path for communicating in a foreign tongue is through instruction. Traditional language pedagogy emphasizes classroom and laboratory practice of vocabulary, grammar, and pronunciation. Lessons are highly structured, with students practicing language skills in class and laboratory. Feedback is offered mostly through exams and drills. However, this classical approach has serious drawbacks, especially when it focuses on declarative knowledge of grammar and vocabulary to the exclusion of conversational skills and comprehension.

Although the ambitious student might achieve conversational fluency by living in a foreign community, this option is unavailable to many. Fortunately, curricula are beginning to incorporate more naturalistic approaches to language learning, powered by technology. The long-term goal is to emulate real-world learning in ways that are effective, economical, and enjoyable.

For computer-assisted language learning (CALL), the "holy grail" is courseware that simulates what a student might experience living in a foreign land. In this virtual community, the student would converse in the target language and receive feedback on ways to improve. Although this pedagogical nirvana won't happen anytime soon, several advances bring it closer to reality. Among these are

(1) powerful, inexpensive computing residing in the "cloud," using a multitude (often thousands) of machines (usually graphical processing units [GPUs]) and abundant memory that mobile devices (e.g., smartphone, tablet) and computers can access easily;
(2) large amounts of online data to "train" pattern classifiers known as artificial neural networks (ANNs);
(3) cloud-based "deep learning" neural networks (DNNs). These are especially powerful ANNs that contain many (often dozens of) hidden layers and intricate connection topologies. A layer is "hidden" if it lies between the input stage of the ANN and its output (i.e., classifier outcome). Each hidden layer adds a level of processing that facilitates the "learning" (through adjustment of activation weights) of features critical for successful classification;
(4) DNN-trained automatic speech recognition (ASR) and speech synthesis (TTS) that deliver a quality and naturalness close to what humans achieve in many (though not all) languages. Many companies (e.g., Amazon, Apple, Google, and Microsoft) use the technology to interact with customers and clients. The data collected are used to further enhance the technology; and

(5) cloud-based virtual and augmented reality applications that extend or replace the user's physical environment through simulation of a variety of situations and environments.

These, along with advances yet to come, will transform the learning experience, not only for language instruction but also for pedagogy in general.

What is the current state of language learning technology, and where is it heading? Before answering, let's first review the history of CALL (Bax, 2003).

A Brief History of Computer-Assisted Language Learning

Computers were introduced into language instruction around 1960 to supplement programmed classroom instruction. Although the technology was primitive by today's standards, early CALL projects demonstrated a potential for enhancing the pedagogical experience. One example is the Programmed Logic for Automatic Teaching Operations (PLATO) Project (University of Illinois at Urbana-Champaign), which included online testing, tutoring, and chat rooms.

Over the years, the quality of CALL improved, driven by advances in interactive media and technology (Warschauer and Healy, 1998). In the 1960s and 1970s, CALL focused on drill and practice lessons in which a computer presented a stimulus and the student responded with (hopefully) the correct response. This was the "structural" (or "restricted") phase of CALL. Beginning in the late 1970s and extending through the early 1990s, CALL entered its "communicative" phase, which emphasized more natural ways of speaking and listening.

With the advent of the World Wide Web and multimedia technology in the 1990s, CALL entered its "integrative" phase, in which the pedagogy was incorporated into a broad range of communication scenarios representative of daily life. During this time, CALL applications offered graphics, animation, audio, and text, all in lessons that combined speaking, listening, reading, and writing (Chapelle and Sauro, 2017).

The key to effective language learning is for the student to use the foreign language as much as possible. Constant practice and feedback is essential. A shortage of language instructors and classroom time makes a compelling case for CALL because it offers instruction anytime, anywhere. Although CALL was originally designed for desktop and laptop computers, its future likely lies with smartphones, tablets, and other mobile devices (e.g., virtual reality [VR] goggles and artificial intelligence [AI]-enabled eyewear).

Computer-Assisted Pronunciation Evaluation and Training

Pronunciation training is where CALL has long deployed cutting-edge technology (Eskenazi, 2009). Several early projects used speech technology to evaluate a student's fluency, pronunciation proficiency, and comprehension. An example is SRI's Autograder project in which Japanese students were evaluated on their ability to speak English intelligibly. An algorithm was developed to emulate intelligibility judgments of native speakers but lacked remedial feedback. Some of this technology was incorporated into PhonePass™, an automatic system for evaluating a student's fluency and proficiency in English (Bernstein and Cheng, 2007).

Both academic (e.g., Carnegie-Mellon, Hong Kong, MIT, Nijmegen, KTH Stockholm) and commercial (e.g., Carnegie Speech, Duolingo, Rosetta Stone®, SRI, Transparent Language®) teams have developed technology that evaluates pronunciation using methods adopted from ASR. At first glance, ASR appears a perfect match for CALL. In place of a language teacher, why not leave the tedium of tutoring to an algorithm embedded in the cloud? It's available 24/7, never tires or sickens, and doesn't go on vacation. However, ASR-based CALL has its drawbacks. For one, ASR doesn't classify individual speech sounds with great precision (e.g., Greenberg and Chang, 2000). Like humans, automatic systems don't decode speech sound by sound but rather rely on clever engineering to infer what the speaker said (or should have said). They do so by culling information from a variety of nonacoustic sources (e.g., location, email, online searches) to supplement the acoustic signal. Although fortuitous for conventional ASR (e.g., Amazon's Alexa, Apple's Siri, and Google Voice), such supplementation can be a serious drawback for CALL applications. This is due to the uncertainty surrounding the identity of specific speech sounds (a.k.a. "phonetic segments" or "phones"). To better understand the problem, let's consider a hypothetical example. The word "pan" consists of three phonetic segments represented by the symbols [p], [æ], and [n] (brackets denote individual segments). An ASR system might correctly identify the initial and final consonants ([p] and [n]) but misidentify the vowel [æ], in which case the word initially "recognized" is "pin" rather than (the spoken word) "pan." However, the vowel's misclassification would probably be overridden by a

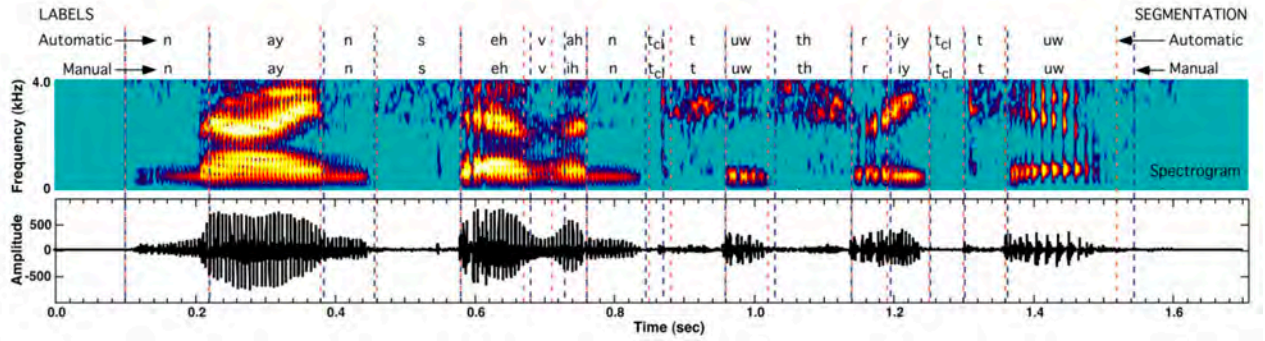

Figure 1. *Alignment of spoken material ("nine," "seven," "two," "three," "two") from the Oregon Graduate Institute "Numbers" corpus (Cole et al., 1994).* ***Top:*** *phonetic labels are similar to those used in the TIMIT corpus (Zue and Seneff, 1988);* ***middle:*** *spectrographic (time versus frequency) representation of the speech signal;* ***bottom:*** *speech pressure waveform. The "automatic" labeling and segment boundaries are analogous to an alignment. The "manual" labels and segment boundaries were provided by a trained phonetician. Reprinted from Chang et al. (2000), with permission.*

"language" model (Chelba and Jelinek, 2000) that factors in semantic context and lexical co-occurrence statistics to identify the word as "pan." ASR-based CALL requires other strategies to compensate for such phonetic imprecision, especially human listener judgments. However, such compensatory methods may themselves compromise the evaluation's accuracy.

An early example of ASR-based CALL was the Voice Interactive Language Training System (VILTS). A student's pronunciation was evaluated by comparing how well the ASR system performed relative to native speakers at a fine-grained level of analysis. The speech signal was partitioned into a sequence of basic sounds ("phones" or "phonetic segments"), and the results of a segment-based ASR system compared. Human raters graded a portion of the student material, and these data served as a baseline for normalizing the ASR-based scores. The system did not offer feedback on how to improve pronunciation (Neumeyer et al., 2000).

VILTS formed the foundation for another SRI system, EduSpeak® (Franco et al., 1999), which evaluates a student's pronunciation. The system comprises several stages: (1) segmentation and labeling (a.k.a. an "alignment") of individual phonetic segments (see **Figure 1**); (2) a measure of the distance between a student's speech and a native-speaker model (based largely on the similarity of their frequency spectra); (3) a comparison of automatically aligned phonetic-segment durations that takes the student's speaking rate into account; and (4) human listener evaluations for calibration. EduSpeak does not require a word transcript but is restricted to languages for which it has been explicitly trained.

Other groups (e.g., Witt and Young, 2000) have also deployed ASR for CALL. Many systems use human listener-based calibration to compensate for the imperfections of ASR. But, as Witt (2012) points out, even human listeners don't necessarily agree on the fine-grained quality of pronunciation (at the segment level), so why should machines be held to a higher standard?

Despite such caveats, several language programs, including Rosetta Stone and Carnegie Speech's NativeAccent®, do offer feedback at the word level (using ASR-based models) that students have found helpful. NativeAccent® also provides rudimentary diagrams of the vocal apparatus as part of its feedback.

Alternatives to ASR-based CALL compare a student's pronunciation to a native speaker's (or rather, a composite model based on a variety of speakers). The more similar the pronunciation of the two, the more intelligible the student's speech is likely to be.

Such a comparison involves both signal processing and acoustic analysis, and includes the following steps.

(1) Phonetic (and other forms of) feature extraction based on a range of spectral and temporal properties for classifying phonetic segments and/or linguistically relevant elements. The most frequent features are (a) a coarse snapshot (25 ms wide) of the acoustic frequency spectrum computed approximately every 10 ms (e.g., Mel Cepstral Frequency Coefficients [MFCCs]; Davis and Mermelstein, 1980); (b) a broadband frequency analysis with relatively fine temporal resolution (a spectrogram as in **Figure 1**); (c) temporal dynamics (velocity ["delta"] and acceleration ["double-delta"] features) of the spectrally filtered speech waveform (Furui, 1986), phonetic-segment and syllable duration as well as the trajectory of the fundamental frequency (pitch contour). A system may also "discover" the most relevant parameters through a process of "feature selection" (e.g., Li et al., 2018).

(2) "Alignment" of a student's speech. This produces a representation of the signal as a sequence of speech sound labels (e.g., [p], [ae], [n]) along with the start and end points of each sound in the speech waveform (these are approximate markers of where a speech sound is likely to begin and end).
(3) A word transcript is often required so that the aligner knows in advance the likely speech sounds and their sequence. The alignment (**Figure 1** shows an example) is based (in part) on acoustic models for each speech sound, often in the context of the sounds that precede and follow. In some systems, such as EduSpeak®, the student's speaking rate (in segments or syllables per second) is estimated as a way of improving the accuracy of the phonetic boundaries.
(4) Dynamic time warping (DTW; Sakoe and Chiba, 1978) is a method for aligning the student's speech, speech sound by speech sound, with a native-speaker composite model. DTW adjusts the segment boundaries to optimize the correspondence between the student and native-speaker model so that a "fair comparison" can be made at the phonetic-segment, syllable, and word levels (**Figure 2**).
(5) A distance metric that quantifies how similar the student's speech is to a native speaker (or rather a composite model comprising many native speakers). The features used for comparison are primarily spectral but may also incorporate dynamic and other temporal properties.
(6) The intrinsic variability of speech, particularly pronunciation, presents a major challenge for CALL technology. To simplify the comparison between the student's utterance and that of a native-speaker model, the analysis recasts the fine-grained spectral and temporal analyses into a form more amenable to quantification.

Because ASR systems have traditionally treated speech as a sequence of short-duration speech sounds (i.e., phonetic segments), it is this analytical framework that is most often used. However, word-level models are becoming increasingly popular and may replace segment models soon.

Three types of pronunciation errors account for most of the pronunciation problems students experience. These mostly occur at the level of individual speech segments (although some problems pertain to syllable prominence and duration). In the discussion that follows, a segment error is underlined to distinguish it from correctly articulated sounds.

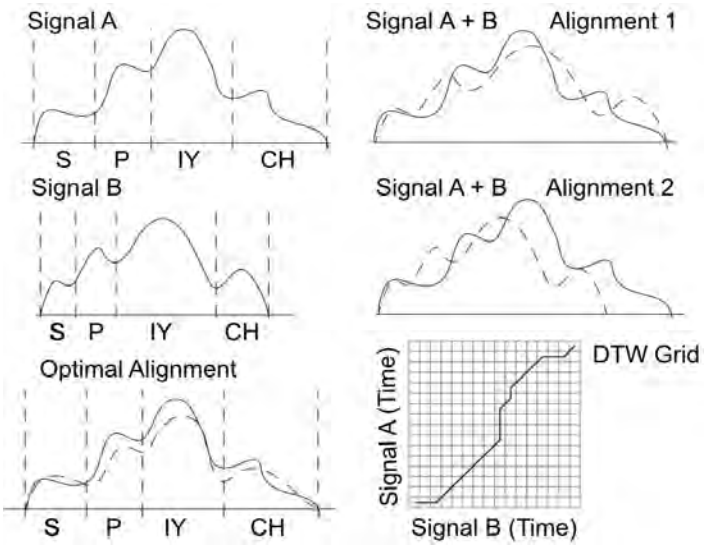

Figure 2. *A highly simplified illustration of dynamic time warping (DTW) to achieve a "fair" comparison of two aligned speech signals. Signals A and B are two instances of the word "speech" spoken by two individuals. Dynamic time alignment iteratively warps the time axis of Signal B until it finds the closest match (in terms of time and spectrum) of the two signals.. The DTW grid shows a hypothetical time warping of Signal B (relative to Signal A) to achieve a quantitatively optimal alignment (i.e., the closest spectrotemporal match) of the two. Signal A and B alignment adapted from Zeng (2000) and DTW Grid adapted from Salvador and Chan (2007), with permission.*

A "substitution" error would be one where the student pronounces the English word "land" as "lend." An "insertion" would occur if the student pronounces "land" as "lands." A "deletion" would occur if "land" were pronounced as "lan" (where the word-final sound [d] is not articulated).

Such departures from the canonical, dictionary pronunciation are one reason why DTW is frequently used to compute the distance between element *X* and element *Y*, where *X* and *Y* may be a word, a phrase, or even longer span of speech (e.g., a sentence).

Because pronunciation is inherently variable (without impacting intelligibility), the distance calculation is usually based on a large number (often hundreds or thousands) of signal parameters. Such complexity is then distilled into a computationally more tractable form using data-reduction methods such as feature selection (e.g., James et al., 2013, p. 203), principal component analysis (e.g., Jolliffe, 2002, or special-purpose neural networks such as autoencoders (Liou et al., 2014). The distance metric may include "high-level" features such as pronunciation error type, intonation, and other pitch properties (e.g., tone level and contour).

Using such comparative methods, Lee and Glass (2015) and Transparent Language's EveryVoice™ technology deploy

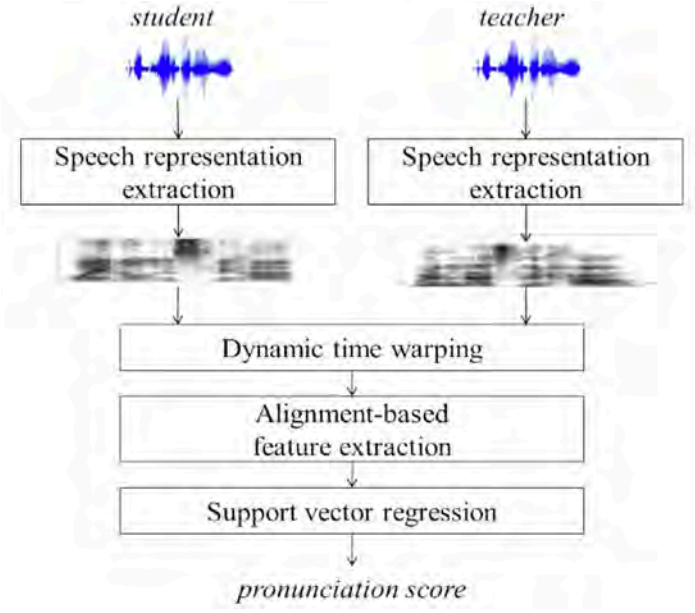

Figure 3. *An early version of Lee's pronunciation evaluation system (see Lee, 2016 for the complete system). After transforming the waveform into a speech representation, the system aligns the two utterances via DTW and then extracts alignment-based features from the aligned path and the distance matrix. A support vector regression analysis (a form of optimization) is used for predicting an overall pronunciation score. Reprinted from Lee and Glass (2013), with permission.*

DTW and DNNs to pinpoint pronunciation errors and offer remedial feedback. Lee's (2016) study is especially instructive. It uses DTW, alignment of the student's speech with a native speaker model, as well as machine learning to flag mispronunciations (**Figure 3** shows a simplified version of their system).

The more similar a student's speech is to the native-speaker model, the more successful the DTW-based alignment is likely to be. Instances where the alignment falters or shows anomalies are flagged as potential mispronunciations. The system also ascertains the specific form of error (substitution, deletion, or insertion). The benefit of this approach is its simplicity and adaptability to a broad variety of languages without the need for extensive customization.

Deep Linguistic Analysis

In classical ASR, a neural network is trained to recognize each of the dozens of different speech sounds (phones) in the phonological inventory of a language. Using a dictionary lookup, it's theoretically possible to represent thousands of words using just a few dozen symbols (which partially overlap with the 26 characters of the Roman alphabet). However, several dozen is still a large number with which to train a neural network, particularly when preceding and following phonetic contexts are considered. Such context-dependent models (Yu and Li, 2015) can number in the thousands, making neural network training especially challenging, especially for limited amounts of training material.

The training of the neural network comes in three basic forms (Bishop, 2006): (1) "supervised," in which the training data are explicitly labeled (as in **Figure 1**), preferably with time boundaries for segments and words; (2) "unsupervised," in which none of the training material is labeled; and (3) "semisupervised," in which a portion of the training is supervised and used for training on unlabeled data.

For many years, supervised training was the norm despite it being burdensome and expensive. Unsupervised and semisupervised training are becoming more popular as DNNs increase in sophistication.

Deep Learning Neural Networks

The architectures of deep neural networks are more complex (and powerful) than classical ANNs as the result of their enhanced connectivity across time and (acoustic) frequency. This power is often augmented with "long short-term memory" (LSTM; Schmidhuber, 2015) and "attention" (Chorowski et al., 2015) models, which further enhance performance. Goodfellow et al. (2016) is an excellent, comprehensive introduction to deep learning and related topics.

Neural networks trained for language instruction have an inherent advantage over those designed for speech dictation and search (e.g., Alexa, Google Voice, Siri) in that the lesson material is often scripted, with most of the words (and their sequence) known in advance. This knowledge makes it somewhat easier to infer which speech sounds have been spoken and in what order (through a pronunciation dictionary). However, this advantage is counterbalanced by the limited amount of data available to train CALL DNNs as well as the diverse (and often unusual) ways students pronounce foreign material.

DNNs have been used in the past where the amount of training material is less than ideal (for classical ANNs). However, even DNNs require a minimum amount of training data to succeed. For this reason, alternative strategies are used to compensate for the paucity of data. A popular approach is to reduce the number of training categories from several dozen to a handful by using a more compact representation such as articulatory-acoustic features (AFs; Stevens, 2002).

What are AFs? They are acoustic models that are based on how speech is produced by the articulatory apparatus. For

Figure 4. *"Pronunciation Practice" in Transparent Language's courseware for Spanish. The student's speech is analyzed and evaluated by deep learning neural network (DNN)-trained acoustic analyzers. Problematic portions are highlighted in* ***yellow****. The student can replay these and compare them with a native speaker. An overall score is shown on the meter (****right****). Published from Transparent Language, with permission.*

example, a phonetic segment can be decomposed into a cluster of acoustic "primitives" that fully distinguish it from other segments in the phonological inventory of the language. Among the most common AFs are "voicing," "manner of articulation," and "place of articulation." Voicing, a binary feature, refers to whether the vocal folds are vibrating (+) or not (−). For example, in the word pan, the initial consonant [p] is "unvoiced," whereas the vowel and following consonant [n] are "voiced." Manner of articulation indicates the mode of articulatory constriction impeding the flow of air through the vocal tract. Examples of manner of articulation categories are "vocalic" (e.g., the vowel in the word "pan"), "stop" (a.k.a. "plosive") consonant (e.g., [p] in "pan"), or nasal consonant (e.g., [n] in "pan"). Place of articulation refers to the locus of maximum vocal tract constriction. In our "pan" example, the initial consonant, [p], has a "bilabial" (both lips) anterior locus of articulation while the final consonant, [n], is produced with the tongue contacting the alveolar ridge (a central place of articulation). Each speech sound (and by extension, syllables and words) can be represented by an analogous set of articulatory features that varies over time.

In EveryVoice™, a native-speaker model based in part on AFs, is compared with the student's utterance by using DTW in concert with several distance metrics. Those speech sounds more than a certain distance from a native-speaker model are highlighted (**Figure 4**). The application also provides an "intelligibility score," which reflects a weighted average of student-native distances across the utterance.

In the future, comparative analyses will likely include a range of time frames (and linguistic levels) such as those associated with the syllable (ca. 200 ms), word (ca. 200-600 ms), and phrase (1-3 s; Greenberg, 1999). Human listeners usually require a second or more of continuous speech to reliably identify all words spoken (Pickett and Pollack, 1963). This extended listening interval is also often required for automatic systems to achieve optimum performance and may account for the recent popularity of "end-to-end" (ETE) and "sequence-to-sequence" (STS) processing in ASR systems (e.g., Prabhavalkar et al., 2017). ETE and STS systems integrate acoustic, pronunciation, and language models into a single, coherent process and so would likely improve the accuracy of ASR-based CALL.

Higher Level CALL Applications

Learning a foreign language involves more than speaking intelligibly. Grammar and vocabulary must also be mastered. Constant practice is key for fluency. Online courseware encourages the student to speak and listen in a broad assortment of realistic situations. In some applications, the student is prompted to respond with a relevant sentence or two. Software evaluates the response, progressing to more difficult material only after the student has demonstrated mastery of the current lesson.

A recent study using ASR goes beyond pronunciation to offer feedback on a variety of language skills, such as grammar and syntax, for students of Dutch (van Doremalen et al.,

2016). A European project, the "Spoken CALL Shared Task" (Baur et al., 2017), offers an illustration of how online evaluation and feedback may operate in the future. The competitive task was based on data collected from a speech-enabled online tool used to help young Swiss German teens practice skills in English conversation. Items were prompt-response pairs, where the prompt is a piece of German text and the response is a recorded English audio file. The task was to "accept" or "reject" responses that may or may not be grammatically and linguistically correct. The task involved more than conventional ASR because it also involved the ability to discern semantically and grammatically appropriate responses using natural language processing. The winning entry (from the University of Manchester, UK) used an ASR system trained with DNNs.

A somewhat different approach is used by the "Virtual Language Tutor" (Wik, 2011), which is an embodied conversational agent that can be spoken to and that, in turn, can talk back (via speech synthesis) to the student. The agent guides, encourages, and provides feedback for mastering a foreign language (initially, Swedish).

The Future of CALL

Several trends in language-learning software are worth noting. Most will likely be enabled through some form of deep learning, among which are the following:

Games

The app FluentU uses real-world video containing music, video, movie trailers, news, and inspiring talks and turns them into personalized language-learning lessons. LingoArcade, Mindsnacks, and DigitalDialects are just a few of the online sites for learning a foreign language using similar material, all within a game-based structure. Su et al. (2013) illustrate several ways to "gamify" dialogue learning for language learning.

Virtual Language Learning

Applications such as ImmerseMe and Mondly place the student in simulated, real-life scenarios, such as a bakery or restaurant, where language skills can be practiced in an engaging way. In these apps, ASR evaluates the student's responses and offers feedback.

Intelligent Language Tutors

Applications such as Duolingo are starting to use "chatbots" to interact with students on a variety of topics to enhance vocabulary and grammar skills. These bots are driven by a combination of ASR, natural language processing, and other forms of artificial intelligence to guide the student through language lessons in naturalistic settings.

Automatic Language Translation

A Defense Advanced Research Projects Agency (DARPA)-funded project, TransTac (Bach et al., 2007), was an early, albeit limited, attempt to provide automatic language translation in a handheld box (for deployment in the Middle East). Among the languages offered were Iraqi Arabic, and Dari. Waverly Labs sells the Pilot™, an earbud-enabled app that performs simultaneous translation in near real time for over a dozen languages. Google Translate offers the ability to translate from one language to another. Among the languages offered for paired translation are English, French, German, Italian, Portuguese, Russian, and Spanish. Google also provides an optical version (using a smartphone camera) that translates signs and other text into one's native language. Microsoft has demonstrated simultaneous translation between English and Mandarin Chinese powered by a DNN that can meld the speaker's voice characteristics with the translated speech. These applications are not especially useful (yet) because they lack the semantic precision and emotional nuance emblematic of human communication, so are best reserved for simple scenarios such as grocery shopping and sightseeing.

Speech Synthesis

The quality and naturalness of speech synthesis has greatly improved, largely due to the ability of DNNs to simulate voices with realism. Baidu's Deep Voice (Arik et al., 2017), Amazon's Polly, Microsoft's Cortana, and Google's Cloud Text-to-Speech (TTS) applications all use DNNs. Google offers TTS in a dozen languages. Deepmind's Wavenet (van den Oord et al., 2017) offers highly realistic synthesis for English and Japanese in multiple voices.

Voice Conversion

Speech synthesis has improved to the point where it is now possible to transform or meld the voice characteristics of one talker into another while preserving intelligibility. Current state-of-the-art systems (Toda et al., 2016) use a special-purpose Vocoder (e.g., STRAIGHT, Kawahara et al., 1999; WORLD, Morise et al., 2016) as the synthesis engine. Two of the more advanced voice conversion systems use DNNs, which include long short-term memory (LSTM)-based recurrent neural networks (Sun et al., 2015) or sequence-to-sequence learning (Miyoshi et al., 2017).

Brain Stimulation

Neurotechnology may play a role in foreign language curricula of the future. A $12 million DARPA grant to Johns Hopkins University (Baltimore, MD) and collaborating institutions explores whether the ability to learn a foreign language can be enhanced through modulating the activation of relevant parts of the auditory and speech areas of the brain through electrical stimulation of the vagus nerve (e.g., Engineer et al., 2015).

Brave New Language-Learning World

DNN-powered speech technology is likely to play an increasingly prominent role in language-learning curricula. As computational power increases and costs diminish, simulation technology will enable a student to inhabit a virtual language world for hours on end. This is likely the future of language instruction, for there is no better way to learn a foreign tongue than to reside in a community where it is spoken. Will it matter that the language community exists only virtually? Virtual reality gaming devices, such as the Oculus Rift™, will only improve over time, enhancing their educational potential. Indeed, language learning could become a "killer app" for educational VR. Stay tuned.

References

Arık, S. O., Chrzanowski, M., Coates, A., Diamos, G., Gibiansky, A., Kang, Y., Li, X., Miller, J., Ng, A., Raiman, J., Sengupta, S., and Shoeybi, M. (2017). Deep voice: Real-time neural text-to-speech. *Proceedings of Machine Learning Research, 34th International Conference on Machine Learning*, Sydney, Australia, August 6-11, 2017, vol. 70, pp. 195-204.

Bach, N., Eck, M., Charoenpornsawat, P., Köhler, T., Stüker, S., Nguyen, T., Hsiao, R., Waibel, A., Vogel, S., Schultz, T., and Black, A. W. (2007) The CMU TransTac 2007 eyes-free and hands-free two-way speech-to-speech translation system. *Proceedings of the International Workshop on Spoken Language Translation* 7.

Baur, C., Chua, C., Gerlach, J., Rayner, M., Russell, M., Strik, H., and Wei, X. (2017) Overview of the 2017 spoken call shared task. *Proceedings of the 7th International Speech Communication Association Workshop on Speech and Language Technology in Education*, Stockholm, Sweden, August 25-26, 2017, pp. 71-78. https://doi.org/10.21437/SLaTE.2017-13.

Bax, S. (2003). CALL—Past, present and future. *System* 31, 13-28.

Bernstein, J., and Cheng, J. (2007). Logic and validation of fully automatic spoken English test. In Holland, M., and Fisher, F. P. (Eds.), *The Path of Speech Technologies in Computer Assisted Language Learning: From Research Toward Practice*. Routledge, Florence, KY, pp. 174-194.

Bishop, C. (2006). *Pattern Recognition and Machine Learning*. Springer-Verlag, New York.

Chang, S., Shastri, L., and Greenberg, S. (2000) Automatic phonetic transcription of spontaneous speech (American English). *Proceedings of the 6th International Conference on Spoken Language Processing*, Beijing, China, October 16-20, 2000, vol. 4, pp. 330-333.

Chapelle, C. A., and Sauro, S. (Eds.). (2017). *The Handbook of Technology and Second Language Teaching and Learning*. Wiley-Blackwell, Hoboken, NJ.

Chelba, C., and Jelinek, F. (2000). Structured language modeling. *Computer Speech and Language* 14, 283-332.

Chorowski, J., Bahdanau, D., Serdyuk, D., Kyunghyun, C., and Bengio, Y. (2015). Attention-based models for speech recognition. *Proceedings of the 28th International Conference on Neural Information Processing Systems*, Montreal, QC, Canada, December 7-12, 2015, vol. 1, pp. 577-585.

Cole, R. A., Fanty, M., Noel, M., and Lander, T. (1994). Telephone speech corpus development at CSLU. *Proceedings of the Third International Conference on Spoken Language Processing (ICSLP1994)*, Yokohama, Japan, September 18-22, 1994, pp. 1815-1818.

Davis, S. B., and Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics Speech and Signal Processing* 28, 357-364.

Engineer, C. T., Engineer, N. D., Riley, J. R., Seale, J. D., and Kilgard, M. P. (2015). Pairing speech sounds with vagus nerve stimulation drives stimulus-specific cortical plasticity. *Brain Stimulation* 8, 637-644.

Eskenazi, M. (2009). An overview of spoken language technology for education. *Speech Communication* 51, 832-844.

Franco, H., Neumeyer, L., Ramos, M., and Bratt, H. (1999). Automatic detection of phone-level mispronunciation for language learning. *Proceedings of the 6th European Conference on Speech Communication and Technology (EUROSPEECH'99)*, Budapest, Hungary, September 5-9, 1999, pp. 851-854.

Furui, S. (1986). On the role of spectral transition for speech perception. *The Journal of the Acoustical Society of America* 80, 1016-1025.

Goodfellow, I., Courville, A., and Bengio, Y. (2016). *Deep Learning*. MIT Press, Cambridge, MA.

Greenberg, S. (1999). Speaking in shorthand — A syllable-centric perspective for understanding pronunciation variation. *Speech Communication* 29, 159-176.

Greenberg, S., and Chang, S. (2000). Linguistic dissection of switchboard-corpus automatic speech recognition systems. *International Speech Communication Association Workshop on Automatic Speech Recognition: Challenges for the New Millennium*, Paris, France, September 18-20, 2000, pp. 195-202.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An Introduction to Statistical Learning*. Springer-Verlag, New York, p. 204.

Jolliffe I. T. (2002). *Principal Component Analysis*, 2nd ed. Springer-Verlag, New York.

Kawahara, H., Masuda-Katsuse, I., and de Cheveigne, A. (1999). Restructuring speech representations using a pitch-adaptive time frequency smoothing and an instantaneous-frequency-based f_0 extraction: Possible role of a repetitive structure in sounds. *Speech Communication* 27, 187-207.

Lee, A. (2016). *Language-Independent Methods for Computer-Assisted Pronunciation Training*. PhD Thesis, Massachusetts Institute of Technology, Cambridge, MA.

Lee, A., and Glass, J. (2013). Pronunciation assessment via a comparison-based system. *Proceedings of Speech and Language Technology in Education (SLaTE 2013)*, Grenoble, France, August 30 to September 1, 2013, pp. 122-126.

Lee, A., and Glass, J. (2015). Mispronunciation detection without nonnative training data. *Proceedings of the 16th Annual Conference of the International Speech Communication Association (Interspeech 2015)*, Dresden, Germany, September 6-10, 2015, pp. 643-647.

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., and Huan, L. (2018) Feature selection: A data perspective. *Association for Computing Machinery Computing Surveys* 50(6), 94.

Liou, C.-Y., Cheng, W.-C., Liou, J.-W., and Liou, D.-R. (2014). Autoencoder for words. *Neurocomputing* 139, 84-96.

Miyoshi, H., Saito, Y., Takamichi, S., and Saruwatari, H. (2017). Voice conversion using sequence-to-sequence learning of context posterior probabilities. *Proceedings of the International Speech Communication Association (Interspeech 2017)*, Stockholm, Sweden, August 20-24, 2017, pp. 1268-1272.

Morise, M., Yokomori, F., and Ozawa, K. (2016) WORLD: A vocoder-based high-quality speech synthesis system for real-time applications. *IEICE Transactions on Information and Systems* 7, 1877-1884.

Neumeyer, L., Franco, H., Digalakis, V., and Weintraub, M. (2000). Automatic scoring of pronunciation quality. *Speech Communication* 30, 83-93.

Pickett, J. M., and Pollack, I. (1963). Intelligibility of excerpts from fluent speech: Effects of rate of utterance and duration of excerpt. *Language and Speech*, 6, 151-164.

Prabhavalkar, R., Rao, K., Sainath, T. N., Li, B., Johnson, L., and Jaitly, N. (2017). A comparison of sequence-to-sequence models for speech recognition. *Proceedings of the International Speech Communication Association (Interspeech 2017)*, Stockholm, Sweden, August 20-24, 2017, pp. 939-943.

Sakoe, H., and Chiba, S. (1978). Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing* 26, 43-49.

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks* 61, 85-117.

Salvador, S., and Chan, P. (2007). Toward accurate dynamic time warping in linear time and space. *Journal of Intelligent Data Analysis* 11, 561-580.

Stevens, K. N. (2002) Toward a model for lexical access based on acoustic landmarks and distinctive features. *The Journal of the Acoustical Society of America* 111, 1872-1891.

Su, P.-H., Wang, Y.-B., Yu, T.-H., and Lee, L.-S. (2013). A dialogue game framework with personalized training using reinforcement learning for computer-assisted language learning. *Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* Vancouver, BC, Canada, May 26-31, 2013, pp. 8213-8217.

Sun, L., Kang, S., Li, K., and Meng, H. (2015). Voice conversion using deep bidirectional long short-term memory based recurrent neural networks. *Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane, QLD, Australia, April 19-24, 2015, pp. 4869-4873.

Toda, T., Chen, L.-H., Saito, D., Villavicencio, F., Wester, M., Wu, Z., Yamagishi, J. (2016). The voice conversion challenge 2016. *Proceedings of the International Speech Communication Association (Interspeech 2016)*, San Francisco, CA, September 8-12, 2016, pp, 1633-1636.

van den Oord, A., Li, Y., Babuschkin, I., Simonyan, K., Vinyals, O., Kavukcuoglu, K., Driessche, G. V. D., Lockhart, E., Cobo, L. C., Stimberg, F., Casagrande, N., Grewe, D., Noury, S., Dieleman, S., Elsen, E., Kalchbrenner, N., Zen, H., Graves, A., King, H., Walters, T., Belov, T., and Hassabis, D. (2017). Parallel WaveNet: Fast high-fidelity speech synthesis. *Proceedings of the 30th International Conference on Neural Information Processing Systems*, Long Beach, CA, December 4-9, 2017.

van Doremalen, J. Boves, L. Colpaert, J., Cucchiarini, C., and Strik, H. (2016). Evaluating automatic speech recognition-based language learning systems: a case study. *Computer Assisted Language Learning* 29, 833-851.

Warschauer, M., and Healey, D. (1998). Computers and language learning: An overview. *Language Teaching* 31, 57-71.

Wik, P. (2011). *The Virtual Language Teacher: Models and Applications for Language Learning Using Embodied Conversational Agents*. Doctoral Dissertation, KTH Royal Institute of Technology, Stockholm, Sweden.

Witt, S. M. (2012). Automatic error detection in pronunciation training: Where we are and where we need to go. *Proceedings of the International Symposium on the Automatic Detection of Errors in Pronunciation Training*, Stockholm, Sweden, June 6-8. 2012, pp. 1-8.

Witt, S. M., and Young, S. J. (2000). Phone-level pronunciation scoring and assessment for interactive language learning. *Speech Communication* 30, 98-108.

Yu, D., and Li, D. (2015). *Automatic Speech Recognition: A Deep Learning Approach*. Springer-Verlag, London.

Zeng, Y. (2000) *Dynamic Time Warping Digit Recognizer.* MS Thesis, University of Mississippi, Oxford.

Zue, V. W., and Seneff, S. (1988). Transcription and alignment of the TIMIT database. *Recent Research Towards Advanced Man-Machine Interface Through Spoken Language*, pp. 515-525.

BioSketch

Steven Greenberg worked on SRI's Autograder project in the early 1990s. More recently, he has collaborated on the development of Transparent Language's EveryVoice™ technology. He has been a visiting professor in the Center for Applied Hearing Research at the Technical University of Denmark, Kongens Lyngby, as well as a senior scientist and research faculty at the International Computer Science Institute in Berkeley, CA. He was a research professor in the Department of Neurophysiology, University of Wisconsin, Madison, and headed a speech laboratory in the Department of Linguistics, University of California-Berkeley. He is president of Silicon Speech, a consulting company based in northern California.