Fully automated speaking assessment: Changes to proficiency testing and the role of pronunciation

Talia Isaacs

UCL Centre for Applied Linguistics, UCL Institute of Education, University College London

Talia.isaacs@ucl.ac.uk

Reference

Isaacs, T. (2018). Fully automated speaking assessment: Changes to proficiency testing and the role of pronunciation. In O. Kang, R. I. Thomson, & J. Murphy (Eds.), *The Routledge handbook of contemporary English pronunciation* (pp. 570–584). Abingdon, UK: Routledge

Abstract

Technological innovation has revolutionized capability in assessing speech, making it possible to record, quantify, and score oral performances outside of live testing conditions. Technology has also increased test-user choice, with fully automated tests, which were initially trained on human ratings but devoid of a human component in test delivery and scoring, now available on the market. The goal of this chapter is to underscore the innovations, trade-offs, and debates in the machine scoring of speech, highlighting the central role of pronunciation. The consequences of machine-driven assessment and possibility of complementing machine scoring with human ratings will also be considered.

Main text

Introduction

Technology has revolutionized capability in assessing speaking, making it possible to record, quantify, and score test-takers' oral performances outside of live testing conditions. Technology has also opened the door to new possibilities in terms of the delivery and scoring of tests, resulting in greater choices for test-users in the second language (L2) speaking assessments available on the market in the 21st century. The focus of this chapter is on fully automated assessments of L2 speech—that is, operational test delivery and scoring of speech production without any human intervention, achieved through both automatic speech recognition (ASR) technology, and an algorithmic score generator to optimize approximation to human ratings. As a preface to discussing how technology has transformed the L2 speaking assessment terrain and the centrality of pronunciation (particularly segmental features) to the state-of-the-art in machine scoring, it is important to situate automated assessment in its broader historical context. After outlining how age-old concerns about the lack of objectivity in human ratings of L2 speech have been addressed through the use of technology, the chapter will focus on operational machine scoring systems used for L2 speaking assessments, including feedback on segmental errors for low-stakes (e.g., research or pedagogical) purposes. The discussion will be underpinned by debates on trade-offs in using fully automated assessments, highlighting concerns about technological constraints dictating the nature of the assessment. Future directions will then be discussed, one of which is ways that technology can be used to prioritize the factors that are most important for intelligibility (i.e., being easily understandable to listeners), instead of simply promoting accent reduction (i.e., mastery of all L2 pronunciation features when some are more consequential for communication than others). Finally, the chapter will conclude with the possibility of using hybrid machine-human scoring systems.

Historical and current conceptualizations of the topic

The L2 assessment literature has long distinguished between three modes of speaking assessment that are foundational to a discussion of the topic. A ‘direct test’ refers to a spoken assessment conducted face-to-face with an interlocutor (e.g., interviewer, examiner, test-taker), whereas a ‘semi-direct test’ is machine administered (e.g., computer- or phone-based), with the test-taker speaking into a recording device. Finally, an ‘indirect test’ denotes assessing speaking without eliciting any form of spoken production (Ginther, 2012). An example of indirect test item is the following multiple-choice question:

Which of the following words is unlike the others in terms of how the underlined sound is pronounced?

(a) treasure *(b) physics* *(c) casual* *(d) lesion*

All words in the example include /ʒ/ except for choice ‘b’, which is pronounced [z]. This discrete-point item is modelled on a prototype item proposed by Lado (1961), an exponent of structural testing, in the most comprehensive practical guide to constructing, administrating, and scoring L2 pronunciation tests that exists today (Isaacs, 2014). Lado’s contention that indirect testing could be used as an alternative to direct or semi-direct testing was predicated on his assumption that written responses of the type exemplified above would strongly correlate with test-takers’ actual L2 pronunciation productions. However, subsequent research confirmed that indirect testing bears little if any relation to test-takers’ oral outputs on the same task, and the statistical associations were even weaker with sentence- and discourse-level speaking tasks (Buck, 1989). Indirect testing is currently not regarded as a valid form of L2 speaking assessment because it strays from the seemingly common-sense assertion that “the best way of assessing how well a learner speaks a language is to get him or her to speak” (He & Young, 1998, p. 1).

Although this major reservation should not be overlooked, indirect testing of speech is attractive for at least two reasons that draw parallels with modern arguments promoting fully automated assessments. First, as Lundeberg (1929) stated, individual oral assessment is “cumbersome and time-consuming” (p. 195), particularly with large numbers of test-takers—a point that motivated Lado’s (1961) suggestion for using indirect tests. To elaborate, human scoring often requires employing and training teachers or examiners to elicit and rate oral productions, which can be resource-intensive in terms of costs and time. Like indirect testing, fully automated assessment is a cheaper option and can generate results more quickly. For example, in English proficiency testing for university entrance purposes, the direct speaking component of the IELTS (British Council, 2016) and semi-direct TOEFL iBT (ETS, 2016) mail score reports to test-takers within 13 calendar days of taking the test, with the TOEFL results available on-line three days earlier. By comparison, the fully automated (computer-mediated and scored) Pearson Test of English (PTE) Academic provides on-line results within five business days (Pearson, 2014). In fact, the automated scoring of a sentence-level speech sample takes only a few seconds for the machine to compute, but on-line file transmission from the test location to the e-scoring site can be subject to broadband issues and other technical challenges (Van Moere & Downey, 2016). Test-takers’ testimonials on Pearson’s website praise the rapid delivery of PTE Academic results within 24 or 48 hours of taking the test, revealing Pearson’s use of quick processing time for marketing purposes.

However, initially developing an ASR and machine scoring system, depending on the purpose and stakes of the assessment, can be large-scale and costly. Sentence-level models that involve machine training to recognize a particular fixed sentence that corresponds to test-takers’ likely spoken output, as elicited through a highly controlled task (e.g., sentence repetition), are much more expensive to develop than phoneme-level models (Bernstein, Cohen, Murveit, Rtischev, & Weintraub, 1990). Both models use phonemes as the “building block,” sequencing sounds based on the probability of occurrence associated with each sound in an effort to decode speech from the soundwave as recognizable output (Franco et al., 2010). The distinction is that phoneme-level models, unlike sentence-level ones, are not tied to particular pre-determined utterances and, thus, can be used with random or impromptu test-taker responses. Sentence-level models, however, are superior in aligning detected phones with the speech signal and approximating human ratings. Once the ASR system has been developed, the machine scoring system is calibrated using a large database of human ratings of a large volume of L2 test-takers’ speech (e.g., thousands of responses produced by hundreds of test-takers in developing Pearson’s automated scoring technology for English; Pearson, 2012), which will inevitably involve some expense. However, regardless of the specifics of the methodology used for test development, wide-scale implementation of the eventual test is likely to more than offset development costs, making it an attractive option for assessment organizations, who may choose to pass on cost savings to test-users.

Another parallel between indirect speaking tests and automated assessment relates to concerns about the inherent subjectivity of human scoring, in part due to the intangible and transient nature of speech (Lundeberg, 1929). For example, the dichotomous task of evaluating whether English learners from a mixed L1 background produce /i/ and /ɪ/ in a target-like or untarget-like way is unlikely to always yield exact rater agreement. Nor is having raters assign each token to one of the two vowel categories that represent the closest target-like approximation (e.g., Thomson & Isaacs, 2009). As Lado (1961) underscored, intelligibility (however measured) is fraught by the issue of intelligible to whom. Finally, global ratings of pronunciation-relevant constructs (e.g., degree of foreign accent), although possibly yielding high interrater reliability, clearly lack the objectivity of scoring multiple-choice items. Fully automated scoring offers a solution to this subjectivity issue that Lado found so pressing, offering standardized delivery and reliable and objective scoring of L2 speech. In so doing, it allows the test developer to draw on some of the advantages of indirect testing, which are disqualifying due to concerns about validity, without actually resorting to indirect testing. Instead, it allows for a semi-direct, automatically scored speaking test.

One advantage of automated scoring that ties into the objectivity argument is that it is impervious to the individual rater idiosyncrasies that are present in human ratings (Xi, 2012). That is, automated scoring eliminates sources of construct-irrelevant variance, which refers to variables extraneous to the L2 speaking ability being measured that could come into play in human ratings (Messick, 1990). For example, individual differences in rater familiarity with or attitudes toward test-takers' L1 accent could unduly influence their speech or pronunciation ratings. Machines, which cannot replicate human listening and scoring processes, are not susceptible to such rater effects (Van Moere & Downey, 2016). However, there are trade-offs to the reliability and efficiency that automation entails, which surface in the next section and permeate the rest of this chapter.

Limitations of ASR and automatic scoring

There are several challenges in undertaking ASR and fully automated scoring of L2 speech that need to be considered before critically evaluating the use of this technology for L2 assessment purposes. First, automated scoring of the spoken medium is much more challenging to implement than automated scoring of writing. This is mainly due to the first step and complicating factor of needing to recognize words from the speech signal (Bridgeman, Powers, Stone, & Mollaun, 2011). Second, word recognition tends to be much better on highly controlled tasks that constrain test-takers' output than on extemporaneous speech tasks that yield relatively unpredictable test-taker output. This is because when the test-takers' speech is predictable or known, the audio track can be forcibly aligned with the transcription of the utterance using pattern matching, with deviations between the expected and actual response more easily detected (Zechner, Higgins, Xi, & Williamson, 2009). Compared to spontaneous speech tasks, automatic scoring of controlled or guided tasks (e.g., sentence read-alouds) result in much stronger correlations between machine scoring and human ratings. Although such tasks may be less authentic in light of the tasks that test-takers need to perform in real-world settings, strong statistical associations with human-rated measures are sometimes used by testing organizations as part of the validity argument to justify their use for high-stakes purposes (e.g., Bernstein, Van Moere, & Cheng, 2010).

A third major challenge lies in decoding L2 learners' as opposed to native speakers' (NS') utterances, particularly when test-takers are from mixed L1 backgrounds due to L1 influence (Isaacs, 2014). Determining a standard that the speech will be compared to, which often begins with a NS corpus, is premised on the notion that L2 speech differs from L1 speech in the pronunciation and sequencing of sounds and words. These deviations are treated as errors, and parameters need to be set so that the system can classify errors and provide learners with feedback on their production accuracy (Cucchiarini, Neri, de Wet, & Strik, 2007; Cucchiarini & Strik, this volume). However, state-of-the-art error detection algorithms are imperfect and occasionally provide learners with misleading feedback about the correctness of utterances (Eskenazi, 2009). More specifically, automatic detection systems for L2 learners sometimes produce false positives (i.e., the system scores the production of a correctly pronounced phone as an error) and false negatives (i.e., the system fails to detect an incorrect phone when one is uttered). Nonetheless, there is some evidence that learners in laboratory contexts can improve their production of targeted phonemes when they receive ASR-based feedback compared to learners who receive no such feedback (Cucchiarini, Neri, & Strik, 2009), highlighting the potential of this technology for targeting segmental accuracy. Unfortunately, not all fully automated tests publish information on the error classification accuracy of their patented technologies, making it difficult to evaluate these systems (e.g., Wagner & Kunnan, 2015).

One remaining challenge and limitation in light of technological capability is that current automatically scored speaking assessments heavily rely on spectral (i.e., frequency-based) and durational (i.e., time-based) measures associated with segmental accuracy and temporal fluency. This is mostly at the exclusion of other measures, since machines are less adept at scoring higher-order features of L2 speaking performances, such as discourse organization, lexical resource, grammatical accuracy, prosody, and content development (Bridgeman et al., 2011). For example, temporal measures (e.g., speech rate, silent pause duration per word), which the machine can ably automatically compute, fail to capture broader notions of fluency that humans may heed in scoring discourse-level tasks, such as coherence and argumentation (Ginther, Dimova, & Yang, 2010). Thus, some aspects of the L2 speaking construct that test developers and examiners may value and take into account when scoring are currently not captured in automated scoring systems (Xi, Higgins, Zechner, & Williamson, 2012). Before discussing the implications of these challenges and constraints, the next section will describe the basic operational processes involved in fully automated scoring by drawing on examples of L2 assessments that use the technology.

Illustrations and examples

Operational ASR and automated feedback and scoring systems

The innovation of fully automated L2 proficiency assessments has resulted in major shifts in the L2 assessment and research terrain in the 21st century. One effect is that the centrality of pronunciation in ASR and automated scoring has, in part, fuelled a resurgence of interest in pronunciation within the L2 assessment community following a lengthy period of neglect (Isaacs, 2014). In order to situate fully automated assessments within the L2 testing context more broadly, the 2x2 matrix in Figure 1.1 provides examples of stand-alone L2 speaking tests, speaking components of L2 proficiency tests, or patented scoring systems (e.g., SpeechRater). Quadrant II shows direct L2 speaking tests or speaking components that are also human scored, Quadrant I shows semi-direct tests that are human scored, and Quadrant IV shows machine-scored semi-direct tests (see also Galaczi, 2010). The tests listed in Quadrant IV are fully automated and, hence, the focus of this review. Notably, no tests have, as yet, been developed in Quadrant III. Producing tests that are human administered and machine scored would remove the advantage of having a more authentic interaction with a human interlocutor on an extemporaneous L2 speech task, as the type of task that could be used would need to be constrained considerably to make it easily machine scorable. The constraints in the design of such a test, the lack of standardization if different human raters administered even a scripted test compared to machine administration, and the lack of a practical advantage entailed in a human needing to test individually would seem to make it a less attractive and perhaps less intuitive option for test developers. On the other hand, it could be argued that test-takers in classroom contexts might prefer speaking tests (however inauthentic and regardless of how it is scored) to be administered by their L2 teacher, although this has yet to be established. An extension of this argument is that, in foreign language contexts that do not routinely teach L2 speaking (Wall & Horák, 2006), integrating speaking into the classroom, even in a way that is reminiscent of audiolingual (as opposed to communicative language) teaching, is better than no speaking focus at all. What is clear is that many more speaking assessments that use automated scoring will be developed in the future. It is important to note, however, that attractive, user-friendly, accessible applications do not mean that fully automated assessments are effective L2 learning tools, nor that they provide valid information about L2 speaking ability, whether the feedback is formal or informal. Automated tests, and particularly those intended for high-stakes purposes (i.e., consequential decision-making) need to be held to the same standard of providing robust validation evidence as traditional human scored tests (see Wagner & Kunnan, 2015).

		L2 test mode of delivery	
		Human	Machine
L2 test scoring	Human	II IELTS Cambridge English exams (e.g., CAE) Oral Proficiency Interview (OPI)	I TOEFL iBT Aptis SOPI, COPI (tape- or computer-based OPI)
	Machine	III NONE	IV PTE (Academic, General) Versant tests ETS's SpeechRater

Figure 1.1. Matrix of L2 speaking assessment administration and scoring (i.e., human- or machine-mediated), with examples in each category

But what about the architecture of assessment-oriented ASR and scoring systems? They tend to consist of an automatic telephone response system (if the test is telephone-mediated; e.g., Versant), a digital storage bank for the recordings, a speech recognizer for decoding the speech, a speech analyser or computation model, which prepares the recognized speech for scoring using mathematical representations, and a score generator, which maps a selection of measures onto an L2 speaking score based on previous ratings by human raters (Pearson, 2011; Zechner et al., 2009). The speech samples on which the system is trained should be produced using professional quality microphones in environments with little ambient noise and recorded with an adequate sampling rate (e.g., 8 kHz for telephonic systems; 16 kHz for computer-mediated systems) and resolution (e.g., 16-bit) to ensure reasonable sound quality. Where possible, operational testing conditions should also avoid noisy environments, provide suitable microphones, use the same sampling rate as was used to generate the acoustic model for audio recording, and conduct sound checks prior to the assessment to ensure adequate volume levels (Chengalvarayan, 2009).

It is practical to illustrate how the technology works by drawing on some examples. EduSpeak, a pronunciation focused computer-assisted language learning (CALL) application (Franco et al., 2010), was reportedly developed using a similar approach to the patented technology used for Pearson's automated speaking tests, providing some information about the development of the ASR technology than is not currently available in Pearson's validation manuals (e.g., Pearson, 2012). Franco et al. (2010) summarize three steps in scoring pronunciation on sentence reading or repetition tasks:

- phonetic segmentation (i.e., identifying the boundaries between phones) using the speech recognizer
- machine generation of scores by comparing a learner's speech to the training database using probabilistic models
- calibration of the scores using an algorithm to combine automatic measures that best map onto scores assigned by human listeners from the training database

As with other ASR systems, EduSpeak was developed using statistical modelling that derives probabilities of sounds occurring within words and, in turn, within sentences. A hierarchical structure is imposed, with sentences modelled as word sequences which, in turn, are modelled phone sequences that vary in their acoustic realization according to their phonetic context. The system combines data trained on both NSs and nonnative speakers (NNSs) of North American English by assigning them different statistical weightings. Scores are computed by averaging the automatically derived scores for each phone in that sentence based on the likelihood of occurrence. Other computed measures are fluency-related (e.g., speech rate, length of phones uttered), normalized over the duration of the sentence or over all of the speaker’s utterances. These measures are then compared to measures from the NS corpus to estimate probability distributions. Final scores for reporting purposes are assigned using statistical modelling to classify the set of scores obtained from the automated system into ratings that approximate predicted human scoring based on the training corpus. Information on the machine’s error rates for phone-level mispronunciation is reported for each phone by comparing the machine scoring with the same speech manually transcribed by human listeners schooled in phonetics (Franco et al., 2010). For comparison purposes, in terms of sentence repetition scoring on Pearson’s Versant English Test (2011), word error frequency is operationalized as the smallest number of segmental substitutions, deletions, or insertions required to modify the test-taker’s output to find the best string match. An utterance that matches the referent speech is scored as error-free, with dysfluencies disregarded in scoring this task type. However, machine misclassification rates (e.g., of individual phones or aggregate measures) are not reported in Pearson’s test validation manuals, suggesting a degree of underreporting.

It is worthwhile discussing Pearson’s Versant English Test (originally developed as PhonePass in the 1990s) in some depth, since this was the first fully automated L2 speaking test available in the high-stakes assessment market, revolutionizing the field (Isaacs & Harding, in press). Versant tests, which are telephone- or computer-delivered (duration: 15–20 minutes,) are currently available in six languages in addition to an aviation English test (Bernstein, et al., 2010). Drawing on Levelt’s (1989) Speech production model, the Versant English Test validation manual defines the speaking construct as the “facility” or “ability to understand spoken English on everyday topics and to respond appropriately at a native-like conversational pace in intelligible English” or, stated differently, the “ease and immediacy in understanding and producing appropriate conversational English" (Pearson, 2011, p. 8). Test items were initially developed for American English, were subsequently sent to NSs of other Inner Circle varieties for checking (e.g., British and Australian English), and were retained only when over 90% of educated NSs responded appropriately to them. Five of six Versant item types are currently operational for machine scoring: reading, repeating, and unscrambling short sentences, providing word length responses to short answer questions, and retelling a story. The other item type, which elicits an opinion or explanation about a defined topic, is, as yet, too unpredictable for the machine to cope with in a way that sufficiently replicates the numerical ratings that humans assign.

The Versant targets "phonological fluency, sentence construction and comprehension, passive and active vocabulary use, listening skill, and pronunciation of rhythmic and segmental units" (p. 3). Score reporting includes four 5-point analytic subscores that, together, comprise test-takers' "facility" reported in an Overall score (p. 18). The Sentence Mastery (syntactic processing) and Vocabulary (lexical comprehension and production) subscales are derived based on the presence of expected lexical items in the expected order. The Fluency subscale captures rhythm, phrasing, and timing (e.g., response latency, speech rate, stops and starts), whereas the Pronunciation subscale refers to "the ability to produce consonants, vowels, and stress in a native-like manner" at the sentence level using everyday vocabulary (p. 12). What is said, which relates to the first two measures, and how it is said, which relates to the latter two, each account for half of the total score, although different subscales are used depending on the task.

Fluency and Pronunciation subscores, which are strongly correlated with the total score (.88 and .86, respectively), are also highly intercorrelated (.80), suggesting that they are measuring a similar construct, compared to much lower intercorrelations between Pronunciation and both Sentence Mastery (.55) and Vocabulary (.51). Correlations between human and machine scoring are lower for Pronunciation (.88) than for the other and overall subscores (.94–.97), but still acceptably high. In terms of concurrent validity, correlations with TOEFL iBT speaking and IELTS speaking are around .75 (Pearson). The resulting scores are aligned with the Common European Framework of References for Languages (CEFR) levels (Council of Europe, 2001). However, the CEFR Global scale excludes pronunciation as a criterion from its descriptors (Isaacs, 2014)—a factor which heavily features in Pearson's automated scoring. Presumably for this reason, Versant scores are also related to a scale that the assessment team reportedly adapted from the CEFR Oral interaction scale descriptors (de Jong & Bernstein, 2001), although published details about the adapted scale are scarce.

Word recognition is based on acoustic models solely trained on L2 speech. These are aligned with or compared to an acoustic model trained on NSs from a range of age and native accent varieties and roughly balanced across gender (Bernstein et al., 2010). Conversion to scores occurs by scaling based on the likelihood of the occurrence of the given feature in the NS model built for that specific task. For example, the algorithm scales silent pauses based on the locus of its occurrence and the probability of a NS pausing that long in that environment (e.g., within a clause). Another example is voice onset time (VOT). Aspiration of the [t^h] sound in the word "take" (voiceless aspirated alveolar stop), for instance, would yield a positive voice onset time (long lag VOT), since voicing occurs only after the release of the aspiration. A lack of aspiration [t], which would diverge from most NS norms, would have a voice onset time of roughly zero (short lag VOT). Thus, the test-taker could be penalized based on divergence from NS norms tied to such durational or spectral measurements. Next, using a nonlinear statistical model, parameter weights are assigned to the resulting values that best approximate estimated human ratings of fluency and pronunciation (Bernstein et al., 2010). The PTE Academic score guide, which uses similar tasks as the Versant, states that “the machine does not need to be told what features of the speech are important; the relevant features and their relative contributions are statistically extracted from the massive set of data when the system is optimized to predict human scores” (Pearson, 2012, p. 52). This is how the stochastic system works. However, the precise combination of measures selected in non-linear models to result in the score that best approximates the human ratings is opaque even to the test developers.

Although in traditional ASR scoring and feedback systems, segmental measures dominate, and pronunciation clearly plays a central role in Person’s approach to automated scoring, this is not ubiquitous (Van Moere & Downey, 2016). Educational Testing Service (ETS), a global competitor to Pearson in the English standardized testing industry, models a markedly different approach. For example, SpeechRater 1.0 was developed to help TOEFL iBT applicants with test preparation through informal feedback on their speaking performances. Due to the complex nature of extemporaneous discourse-level speaking tasks on the TOEFL iBT, SpeechRater’s job of recognizing and automatically scoring the speech is much more challenging than would be the case with the highly constrained word- or sentence-level Pearson style tasks (Xi et al., 2012). Using variable sets of test-takers' responses in terms of L2 speaking proficiency level, correlations between trained TOEFL raters’ operational ratings and SpeechRater’s scores were between .65 and .69 (Bridgeman et al., 2011). This is much lower than the correlations obtained for Pearson tests with human raters, as reported above (Pearson, 2011). In several outputs centering on test validation and issues such as construct coverage (e.g., Xi et al., 2012), ETS researchers argue that until SpeechRater’s performance is more optimal, which would necessitate major improvements in the technological capability of the state-of-the-art in ASR and automatic scoring, the intended uses of SpeechRater will remain low-stakes, which they deem appropriate. This is in contrast to Pearson tests, which are currently used for high-stakes purposes (e.g., government visas, academic admissions, professional certification).

Notably, Pearson (2012) does not publish a list of possible machine-generated measures that the algorithm could draw from (there could plausibly be hundreds or thousands) to optimize measurement and automatically derive test scores. In contrast, ETS provides a list of 29 automated speech measures that were candidates for selection for two statistical scoring models explored for use with SpeechRater 1.0. Eleven of these were ultimately selected in a final multiple regression model that was chosen, with fixed weights assigned for different measures as pre-specified by a panel of experts (Zechner et al., 2009). This advisory group took into account factors such as statistical efficiency, approximation to human scoring, and construct coverage, as reflected in the link between the automated measures, and the descriptors used in the TOEFL iBT Speaking Scoring Rubrics (Educational Testing Service, 2009). Seven of the 11 included measures are fluency-related (e.g., mean duration of long pauses; silent pause duration per word) and two are lexical richness measures (types), expressed over the duration of the word or speech sample. The remaining measures are the “language model,” related to grammatical accuracy, and the “acoustic model,” related to segmental errors identified using phoneme sequence probabilities (p. 890). With the exception of the language model measure, all included measures can be related to the descriptors in the TOEFL iBT Delivery subscale, which is one of three subscales contributing to overall TOEFL speaking scores. Conversely, features related to the Language Use subscale (i.e., lexical and grammatical resource) are only partially represented in the scoring model, and none of the measures link to Topic Development (e.g., content relevance, coherence) due to current technological limitations (Zechner et al., 2009). Therefore, although automated scoring eliminates sources of construct-irrelevant variance that play into human ratings, resulting in more objective and efficient scoring, there are trade-offs.

SpeechRater only measures a portion of the related features that make up the multifaceted TOEFL speaking construct, failing to assess the full range of construct-relevant linguistic features reflected in the scales (Bridgeman et al., 2011). It also does not encapsulate the breadth of linguistic properties that human raters reportedly heed in their scoring decisions, including prosody, task completion, and discourse-level features (Brown, Iwashita, & McNamara, 2005). The sections below will refer to this and other practical challenges and ways forward.

New directions and recommendations

This chapter has described technological advances in machine recognition and scoring of speech. These innovations have resulted in new ways of assessing L2 speaking, removing longstanding concerns about subjective element of rating (e.g., Lundeberg, 1929), since the machine, once trained, will always reach the same result. The introduction of fully automated assessment of L2 proficiency has resulted in major shifts in the L2 assessment market, accentuating differences between products (e.g., direct vs. fully automated speaking tests). It is difficult to predict how technology will push the assessment field forward in the generations to come. Perhaps the growing use of interactive spoken dialogue systems will lead to tests involving test-takers creating avatars of themselves and interacting with a virtual teacher or interlocutor, who would be pre-programmed to respond to their spoken responses, providing feedback or leading simulated interactions (Wik & Hjalmarsson, 2009). The possibilities may appear limitless. For example, it may be that ASR and scoring systems will one day move beyond monologic tasks to capture interactional activities involving multiple talkers and overlapping speech, locating interlocutors in time and space. Although machines are able to predict human scoring, they will never be able to replicate human processing. Humans will remain the ultimate arbitrators of whether and the extent to which communication has been successful in real-world settings, and there will always be constraints in what machines are able to do (Isaacs & Harding, in press).

When fully automated assessments intended for high-stakes purposes began to be actively marketed to test-users on the global stage in the first decade of the 21st century, field-wide debates surrounding this development were initially heated and even acerbic. For example, critics from a sociocultural perspective decried the inauthenticity of Versant's decontextualized items (Chun, 2006). However, Pearson (2011) contends that by using tasks that are not contextually embedded, the test excludes cultural schema and social factors from the assessment of L2 speaking ability. They argue that this is more efficient for assessing actual speaking ability by spending the time eliciting test-takers' speech samples rather than creating context. Further, the test correlates with context-dependent speaking tests (e.g., TOEFL, IELTS), supporting the claim that test scores are related to speaking ability on contextually richer speaking tasks (Van Moere & Downey, 2016). Despite these and other rebuttals from the test provider, including about the psycholinguistic nature of the test construct, concerns about validity and authenticity still resonate. In fact, most Versant tasks are more reminiscent of now dated grammar translation and audiolingual style activities than they are of more contemporary style communicative tasks. These points aside, the language assessment community has now transitioned to a more pragmatic understanding that automated assessments are here to stay (Xi, 2010). The next section turns to the topic of setting pedagogical and assessment priorities in relation to features being targeted using automated scoring.

Defining pedagogical priorities

Not all errors are created equal, with some being more detrimental for communication than others (Derwing & Munro, 2015). This chapter has drawn mostly on segmental examples but would not be complete without reference to prosody. There is growing evidence that prosodic features are important for listener understanding of L2 speech (e.g., Kang, Rubin, & Pickering, 2010). However, prosody is difficult for ASR to target. For example, comparing an L2 learner's pitch variation over time to that of referent speakers from a training corpus is more difficult than examining segmental features analyzed segment by segment (Eskenazi, 2009). In addition, prosodic features tend to be amenable to sociolinguistic variables such as age, gender, social class, or geographical variety, making it complicated to determine acceptable deviations from the norm (van Santen, Prud'hommeaux, & Black, 2009). Thus, taking into account inter-speaker L2 prosodic variation, particularly in contexts that allow for different target NS varieties, is challenging. It is relatively easy to set cut-offs in terms of acoustic space for vowel formants (Deng & O'Shaughnessy, 2003), with more or less stringent criteria applied depending on the language varieties accepted and the desired difficulty or leniency of the automated system. However, identifying a similarly narrow range for prosodic features, even after normalizing for vocal tract size differences, is not currently feasible. Van Santen et al. (2009) suggest using prosodic minimal pairs or lexical items with particular stress patterns as tasks to avoid unwanted prosodic variation. However, this would necessitate even more constrained speaking prompts than are currently common for ASR. To summarize, minimal focus on prosody is a limitation of current automated scoring.

In terms of segmental errors, research on the sound contrasts that most impede communication can inform instructional and assessment priorities (Isaacs, 2014) and this extends to automated scoring. Carnegie Speech's commercial NativeAccent program for training purposes detects mispronunciations of phonemes (Pelton, 2012), including some consonant cluster strings and minimal pairs (e.g., /t/ substituted for /θ/) that are unlikely to actually interfere with intelligibility (see Derwing & Munro, 2015). The cautionary note is that linguistic features that are easy for a machine to detect and score may not be of much importance for communication. Therefore, claims by test providers about how the selected features affect intelligibility may be moot, unsubstantiated, or even contradictory of existing evidence. In the case of NativeAccent, the tool's claim to "teach students to speak the language intelligibly in less training time" (Pelton, 2012, p. 11) is likely both overstated and misleading. The element of accent reduction that the software is targeting may be incompatible with helping learners become intelligible (Levis, 2005), with time likely better spent giving learners feedback other pronunciation features.

It is useful to illustrate a related point in reference to a high-stakes automated test. In score reporting for the PTE Academic image (e.g., graph) description task, the proportion of uttered words detected as unintelligible is specified at the three lowest levels of the 6-level pronunciation scale (0-5), with half of the words unintelligible at level 0, a third unintelligible at level 1, and over two-thirds intelligible at level 2 (Pearson, 2012). However, at the scalar extremes, the descriptors of "native-like" (level 5) and "non-English" pronunciation (level 0) suggest that what is being measured is not intelligibility, but rather deviations from what human raters consider to be NS norms, which is reflected in the machine training and scoring methodology (p. 54). The issue here is that being more intelligible does not incrementally increase with sounding more native-like (Derwing & Munro, 2015). In sum, although automated speaking tests may claim to assess intelligibility, most of the emphasis tends to be placed on pronunciation accuracy or congruence with NS norms.

Cucchiarini et al. (2007) demonstrate one approach to preventing an ASR system from indiscriminately targeting all learner errors in an experimental CALL setting. They pre-defining error types as "relevant" (p. 2182) for learners of Dutch from different L1 groups using the following criteria:

- common to speakers from different L1 backgrounds
- perceptually salient

- potentially impede communication
- frequent
- persist over time

After training the model on a NS passage, it was applied to detect segmental (substitution, epenthesis, deletion) errors in addition to using a silent pause model to detect undue dysfluencies. Measurement thresholds for each phone were derived by artificially introducing errors in the NSs' productions and comparing the manipulated speech to the error-free NS productions for the selected targets. The efficacy of model classifications and pre- and post-test results comparing targeted and untargeted sounds were reported.

Raux and Kawahara (2002) describe an alternative approach to pre-specifying segmental errors (target: 10 epenthesis, deletion, and substitution errors) for Japanese learners of English using a probabilistic algorithm to relate intelligibility to error rates detected by their ASR system. A human rater scored learners’ read-aloud productions targeting each error type using a Likert-type intelligibility scale, with the ratings also used to derive an “error priority function” suggesting the most crucial segmental errors for intelligibility (p. 737). They then evaluated the model by eliciting additional data. The model was trained by calculating the error rate distribution of the speech samples for each of the five intelligibility levels, with constraints applied to the formula in light of the assumption that error rates decrease with intelligibility gains (although see Harding, 2017, for problems with this approach). Finally, they examined the difference between the error rate of each learner on the 10 pre-specified errors and the pooled performance of learners at that ability level, with underperformance relative to the mean for features deemed relevant suggesting that it could be a pedagogical focus. Despite the methodological shortcomings of this small-scale study, we can conclude that it is possible to develop an approach to scoring L2 segmental accuracy taking into account presumed pedagogical importance (Eskenazi, 2009).

Is hybrid machine-human scoring a way forward?

One of the issues that has permeated this chapter is that fully automated assessment can result in what Galaczi (2013) has termed a reductionist approach to L2 speaking and a narrowing of the construct. Such tests fail to capture, among other things, different interactional patterns that are typical of tests that adopt an expansive approach to assessing speaking, which is at the other end of the spectrum (e.g., Cambridge First Certificate of English). Bennett and Bejar (1998) articulate the potential risks to validity when constraints of automated scoring having to do with the mode of delivery dictate the way that the assessment is carried out

> *…the interface helps set task and construct parameters. In the worst case, it can unintentionally distort the task and construct definition, either by making the mechanics of response entry so difficult that the responses gathered are not fully reflective of examinee competence or by constraining the substantive aspect of the task to the point that it no longer represents the original construct definition* (p. 11).

One way of mitigating limitations in technology-driven assessment is by having automatic scoring operate in tandem with human ratings. ETS' Test of English-for-Teaching (TEFT), the assessment component of an on-line teacher professional development training program, demonstrates one way of combining machine and human scoring (Zechner et al., 2015). In contrast to the relatively unpredictable discourse-level TOEFL iBT tasks, the TEFT uses controlled tasks more reminiscent of the Versant (Pearson, 2011). To address difficulties that arose in a pilot administration of SpeechRater when it was applied to the test, ETS researchers developed a system whereby test-takers' responses viewed as unscorable by human raters, either due to poor recording quality (e.g., background noise, equipment issues) or to problematic responses (e.g., not in English, off topic), which would have resulted in lower reliability had they been automatically scored, were filtered out and rerouted to a human rater. In this way, all non-problematic items could be machine scored and all screened problematic ones could be human scored. Poorly recorded items raise questions about the validity of the assessment, with intelligibility speech possibly confounded with poor sound quality (Munro, 1998). However, this example demonstrates that, rather than being an either-or option, human and machine scoring could be used in complementary ways in assessing test-takers' performances. A hybrid human-machine scoring system could consist of the machine scoring the elements it does best, including segmental and fluency measures. This way, instead of human raters providing holistic scores on overall speaking ability or on all facets of the performance using detailed analytic scales, an alternative could be implemented. Raters could instead be asked to focus solely on and provide ratings for some elements of the performance not already being assessed (and not scored well, if at all) by the machine (e.g., cohesion, idea development, task execution). This could free up raters' attentional space to concentrate on discrete aspects of speech when rating, potentially simplifying some of the complexity of the rating task (see Lumley, 2005). In sum, one way of offsetting the limitations of technology-mediated assessment and allowing for greater construct coverage is by having automatic scoring complement human ratings—an area for future exploration.

Acknowledgments

Some of the ideas in this chapter arose through an automated speaking assessment project supported by the British Council. I gratefully acknowledge Fred Davidson for the push to explore this area, Judith Fairbairn, Jamie Dunlea, and Barry O'Sullivan for the opportunity to do so, Evelina Galaczi for sharing her thinking on the topic, and the editors of this volume for their helpful comments on this chapter.

References

Bennett, R. E., & Bejar, I. I. (1998). Validity and automad scoring: It's not only the scoring. *Educational Measurement: Issues and Practice, 17*(4), 9-17.

Bernstein, J., Cohen, M., Murveit, H., Rtischev, R., & Weintraub, M. (1990). Automatic evaluation and training in English pronunciation. *Proceedings of the International Conference on Spoken Language Processing* (*ICSLP*) 1990. Kobe, Japan.

Bernstein, J., Van Moere, A., & Cheng, J. (2010). Validating automated speaking tests. *Language Testing, 27*(3), 355–377.

Bridgeman, B., Powers, D., Stone, E., & Mollaun, P. (2011). TOEFL iBT speaking test scores as indicators of communicative language proficiency. *Language Testing, 29*(1), 91–108.

British Council. (2016). *IELTS results process.* Retrieved from July 6, 2016, from http://takeielts.britishcouncil.org/find-out-about-results/results-process

Brown, A., Iwashita, N., & McNamara, T. F. (2005). An examination of rater orientations and test-taker performance on English for academic purposes speaking tasks. Monograph Series MS-29. Princeton, NJ: ETS.

Buck, G. (1989). Written tests of pronunciation: Do they work? *ELT Journal, 43*(1), 50–56.

Chengalvarayan, R. (2009). *Sampling rate independent speech recognition. United States Patent No. 2009/0012785 A1*. Naperville, IL: Patent Application Publication.

Chun, C. W. (2006). An analysis of a language test for employment: The authenticity of the PhonePass test. *Language Assessment Quarterly, 3*(3), 295–306.

Council of Europe. (2001). *Common European Framework of Reference for languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press.

Cucchiarini, C., Neri, A., de Wet, F., & Strik, H. (2007). *ASR-based pronunciation training: scoring accuracy and pedagogical effectiveness of a system for Dutch L2 learners.* Proceedings of Interspeech 2007. Antwerp, Belgium.

Cucchiarini, C., Neri, A., & Strik, H. (2009). Oral proficiency training in Dutch L2: The contribution of ASR-based corrective feedback. *Speech Communication, 51*(10), 853-863.

Cucchiarini, C., & Strik, H. (this volume). Automatic speech recognition for second language pronunciation assessment and training. In O. Kang, R. I. Thomson & J. Murphy (Eds.), *Routledge handbook of contemporary English pronunciation*.

de Jong, J. H. A. L., & Bernstein, J. (2001). Relating PhonePass overall scores to the Council of Europe framework level descriptors. Proceedings of the *European Conference of Speech Communication and Technology* (*EUROSPEECH*) 2001. Aalborg, Denmark.

Deng, L., & O'Shaughnessy, D. (2003). *Speech processing: A dynamic and optimization-oriented approach*. New York: Marcel Dekker.

Derwing, T. M., & Munro, M. J. (2015). *Pronunciation fundamentals: Evidence-based perspectives for L2 teaching and research*. Amsterdam: John Benjamins.

Educational Testing Service. (2009). The official guide to the TOEFL test (3rd ed.). New York: McGraw-Hill.

Eskenazi, M. (2009). An overview of spoken language technology for education. *Speech Communication, 51*(10), 832-844.

ETS. (2016). *Getting your TOEFL iBT test scores.* Retrieved July 6, 2016, from http://www.ets.org/toefl/ibt/scores/get/

Franco, H., Bratt, H., Rossier, R., Gadde, V. R., Shriberg, E., Abrash, V., & Precoda, K. (2010). EduSpeak: A speech recognition and pronunciation scoring toolkit for computer-aided language learning applications. *Language Testing, 27*(3), 401–418.

Galaczi, E. D. (2010). Face-to-face and computer-based assessment of speaking: Challenges and opportunities. In L. Araújo (Ed.), *Proceedings of the Computer-based Assessment (CBA) of Foreign Language Speaking Skills* (pp. 29-51). Brussels: European Commission.

Galaczi, E. D. (2013). Speaking assessment: Evolving and adapting to a changing world *Cambridge English centenary symposium on speaking assessment*. Cambridge: Cambridge English Language Assessment.

Ginther, A. (2013). Assessment of speaking. In C. A. Chapelle (Ed.), *The encyclopedia of applied linguistics*. Hoboken, NJ: Blackwell. Doi: 10.1002/9781405198431.wbeal0052

Ginther, A., Dimova, S., & Yang, R. (2010). Conceptual and empirical relationships between temporal measures of fluency and oral English proficiency with implications for automated scoring. *Language Testing, 27*(3), 379–399.

Harding, L. (2017). What do raters need in a pronunciation scale? The users' view. In T. Isaacs & P. Trofimovich (Eds.), *Second language pronunciation assessment: Interdisciplinary perspectives* (pp. 12–34). Bristol, UK: Multilingual Matters.

He, A. W., & Young, R. (1998). Language proficiency interviews: A discourse approach. In R. Young & A. W. He (Eds.), *Talking and testing: Discourse approaches to the assessment of oral proficiency* (pp. 1-24). Amsterdam: John Benjamins.

Isaacs, T. (2014). Assessing pronunciation. In A. J. Kunnan (Ed.), *The companion to language assessment* (pp. 140–155). Hoboken, NJ: Wiley-Blackwell.

Isaacs, T., & Harding, L. (in press). Research timeline: Pronunciation assessment. *Language Teaching*, *50*(3).

Kang, O., Rubin, D., & Pickering, L. (2010). Suprasegmental measures of accentedness and judgments of language learner proficiency in oral English. *Modern Language Journal, 94*(4), 554–566.

Lado, R. (1961). *Language testing: The construction and use of foreign language tests*. London: Longman.

Levelt, W. J. M. (1989). *Speaking: From intention to articulation*. Cambridge, MA: MIT Press.

Levis, J. (2005). Changing contexts and shifting paradigms in pronunciation teaching. *TESOL Quarterly, 39*(3), 369–377.

Lumley, T. (2005). *Assessing second language writing: The rater's perspective*. Frankfurt: Peter Lang.

Lundeberg, O. K. (1929). Recent developments in audition-speech tests. *The Modern Language Journal, 14*(3), 193-202.

Messick, S. (1990). *Validity of test interpretation and use*. Research Report 90-11. Princeton, NJ: Educational Testing Service.

Munro, M. J. (1998). The effects of noise on the intelligibility of foreign-accented speech. *Studies in Second Language Acquisition, 20*(2), 139-154.

Pearson. (2011). *Versant English Test: Test description & validation summary*. Palo Alto, CA: Pearson Education.

Pearson. (2012). *PTE Academic: Score guide*. n.p: Pearson Education.

Pearson. (2014). PTE Academic: Test centres and fees. Retrieved July 6, 2016, from http://pearsonpte.com/test-takers/test-centres/

Pelton, G. (2012). Mining pronunciation data for consonant cluster problems. Paper presented at the *International Symposium on Automatic Detection of Errors in Pronunciation*

Training (*IS ADEPT*), Stockholm, Sweden. Retrieved July 6, 2016, from: http://www.speech.kth.se/isadept/presentations/Garrett_Pelton.pdf

Raux, A., & Kawahara, T. (2002). *Automatic intelligibility assessment and diagnosis of critical pronunciation errors for computer-assisted pronunciation learning.* Proceedings of the International Conference on Spoken Language Processing (ICSLP) (pp. 737-740). Denver, CO.

Thomson, R. I., & Isaacs, T. (2009). *Within-category variation in L2 English vowel learning*. Proceedings of the annual conference of the Canadian Acoustics Association, Niagra-on-the-Lake, ON (pp. 138–139).

Van Moere, A., & Downey, R. (2016). Technology and artificial intelligence in language assessment. In D. Tsagari & J. Banerjee (Eds.), *Handbook of second language assessment* (pp. 341-358). Berlin: De Gruyter Mouton.

van Santen, J. P. H., Prud'hommeaux, E. T., & Black, L. M. (2009). Automated assessment of prosody production. *Speech Communication, 51*(11), 1082-1097.

Wagner, E., & Kunnan, A. J. (2015). The Duolingo English Test. *Language Assessment Quarterly, 12*(3), 320-331.

Wall, D., & Horák, T. (2006). *The impact of changes in the TOEFL examination on teaching and learning in Central and Eastern Europe: Phase 1, the baseline study*. TOEFL Monograph MS-34. Princeton, NJ: ETS.

Wik, P., & Hjalmarsson, A. (2009). Embodied conversational agents in computer assisted language learning. *Speech Communication, 51*(10), 1024-1037.

Xi, X. (2012). Validity and the automated scoring of performance tests. In G. Fulcher & F. Davidson (Eds.), *The Routledge handbook of language testing* (pp. 438-451). Abingdon, UK: Routledge.

Xi, X., Higgins, D., Zechner, K., & Williamson, D. (2012). A comparison of two scoring methods for an automated speech scoring system. *Language Testing*, *29*(3), 371-394.

Zechner, K., Chen, L., Davis, L., Evanini, K., Lee, C. M., Leong, C. W., . . . Yoon, S.-Y. (2015). Automated scoring of speaking tasks in the Test of English-for-Teaching (TEFT). *ETS Research Report Series* (RR15-31). Princeton, NJ: Educational Testing Service.

Zechner, K., Higgins, D., Xi, X., & Williamson, D. M. (2009). Automatic scoring of non-native spontaneous speech in tests of spoken English. *Speech Communication*, *51*(10), 883–895.