



## Speech Technologies and the Assessment of Second Language Speaking: Approaches, Challenges, and Opportunities

Diane Litman, Helmer Strik & Gad S. Lim

To cite this article: Diane Litman, Helmer Strik & Gad S. Lim (2018) Speech Technologies and the Assessment of Second Language Speaking: Approaches, Challenges, and Opportunities, Language Assessment Quarterly, 15:3, 294-309, DOI: [10.1080/15434303.2018.1472265](https://doi.org/10.1080/15434303.2018.1472265)

To link to this article: <https://doi.org/10.1080/15434303.2018.1472265>



Published online: 04 Jun 2018.



Submit your article to this journal [↗](#)



Article views: 1272




View related articles [↗](#)



View Crossmark data [↗](#)



# Speech Technologies and the Assessment of Second Language Speaking: Approaches, Challenges, and Opportunities

Diane Litman<sup>a</sup>, Helmer Strik<sup>b</sup>, and Gad S. Lim <sup>c</sup>

<sup>a</sup>Department of Computer Science, University of Pittsburgh, Pittsburgh, PA, USA; <sup>b</sup>Centre for Language Studies, Radboud Universiteit, Nijmegen, Netherlands; <sup>c</sup>Michigan Language Assessment, Ann Arbor, MI, USA

## ABSTRACT

This article provides an overview and evaluation of the uses—actual and potential—of automatic speech recognition (ASR) and spoken dialogue systems (SDS), related technologies that can be applied to second language speaking assessment, given particular definitions of the construct. Both technologies have only gradually moved in the direction of supporting language learning, and only more recently used for grading purposes. How the speaking construct is defined determines one's evaluation of the extent to which assessments using these technologies are adequate to the task, given different test use contexts, and what the challenges and future research requirements are. In any event there are many opportunities for their use in assessment, and these would be facilitated by increased cross-disciplinary research among the language testing and speech technology communities.

## Introduction

In considering what second language speaking assessment might look like in the future, it would be negligent not to consider the role that speech technologies might have, given their increasing ubiquity in many aspects of everyday life. Speech technologies can potentially be used both in the delivery and scoring of speaking assessments, providing that the requirements of validity can be met. How close these technologies are to being ready depends, among other things, on the construct being tested.

The other articles in this special issue reflect different views of the speaking construct—from the psycholinguistic (De Jong, [this issue](#)) to the communicative (Galaczi & Taylor, *this issue*)—and different technologies are needed to implement assessments embodying different constructs. From a psycholinguistic perspective, processing efficiency (e.g., response time, rate of delivery, absence of pauses) can be evidence of speaking ability (Van Moere, 2012). Thus, what is required of technology is capturing some speech sample in a way where those variables can be recognized and measured. On the other hand, where the construct includes notions such as interactional competence, then technology with dialogue capabilities would also be required.

With the above in mind, this article reviews the past, present, and potential future use of two key and related types of technology for delivering automated speech assessments covering the range of constructs discussed—namely, automatic speech recognition (ASR) and interactive spoken dialogue systems (SDS) that are built on top of ASR systems—focusing on the research challenges as well as opportunities in this area.

## Automatic speech recognition

The goal of a standard ASR system is to determine which spoken words are present in a given speech (audio) signal. Such ASR systems generally consist of a decoder (the search algorithm) and three “knowledge sources”: the lexicon, the language model, and the acoustic models (Figure 1). An ASR system will first have to be trained. The input consists of the lexicon and a large corpus of audio files, which are used to produce the language model and the acoustic models. The language model contains probabilities of words and sequences of words. Acoustic models are models of how the sounds of a language are pronounced. The lexicon is the connection between the language model and the acoustic models. It therefore contains two representations for every entry: an orthographic and a phonological transcription. Because words can be pronounced in different ways, lexicons often contain more than one entry for some words (i.e., the pronunciation variants). After a system has been trained, new speech samples can be presented to the system, which will then use the different knowledge sources to recognize what was said.

### History of ASR

The history of ASR research goes back to the Defense Advanced Research Projects Agency (DARPA) Speech Understanding Research project, which was carried out at different sites and run from 1971 to 1976. Technologies from this period, such as hidden Markov modeling (HMM) and dynamic programming, remain in use today. Over time, there have been significant improvements in the way ASR systems in general come to recognize speech. Apart from HMMs, artificial neural networks (ANNs) and hybrid HMM-ANN systems have been used. More recently, it has been shown that ASR performance can be increased substantially by using deep neural networks (DNNs) for ASR (e.g., see Pereyra, Tucker, Chorowski, Kaiser, & Hinton, 2016; Xiong *et al.*, 2016).

The history of ASR for learner speech is relatively more recent, starting in the late 1990s. At the CALICO conference in 1996 a “special interest group” on Computer Assisted Pronunciation Investigation Teaching And Learning (CAPITAL) was formed, and in 1998, the workshop Speech Technology in Language Learning (STiLL) was organized in Marholmen, Sweden. STiLL was followed by three InSTiLL symposia: in Besançon (Delcloque, 1999), Dundee (Delcloque, 2000), and Venice (Delmonte, 2004). At Interspeech-

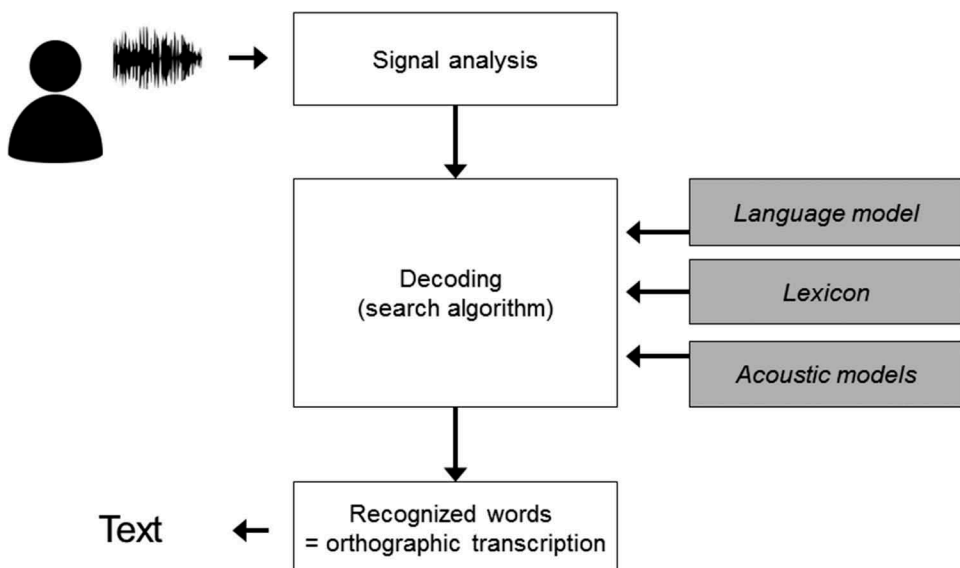


Figure 1. A prototypical automated speech recognition system.

2006 in Pittsburgh, there was a special session on Speech and Language in Education (URL-IS06), and the Speech and Language Technology in Education (SLaTE) SIG was started. Since then, there have been seven SLaTE workshops.

Initial attempts to apply ASR to language learning were sometimes met with disappointing results. These were largely the result of using systems developed for native speakers and for other purposes, such as dictation systems, without modification (Coniam, 1999; Derwing, Munro, & Carbonaro, 2000). While ASR of native speech is already complex enough, because of well-known problems, such as background sounds, low signal-to-noise ratio (SNR), disfluencies, pronunciation variation, and the fact that words are not clearly separated in speech, ASR of learner speech is even more complex, because the grammar, the words used, and the pronunciation can deviate considerably from what is expected, thus affecting all three “knowledge sources” of the ASR system. Furthermore, speech in a second language (L2) tends to contain more disfluencies and hesitation phenomena than native speech, which can manifest in different ways, depending on learners’ proficiency level (Cucchiariini, Van Doremalen, & Strik, 2010). These differences between L1 and L2 speech are so extensive that the ASR performance is degraded considerably (Van Compernelle, 2001; Van Doremalen, Cucchiariini, & Strik, 2010, 2011). Consequently, the negative findings obtained in the early days relate to the inappropriate use of those systems rather than to ASR technology in general (Strik, Neri, & Cucchiariini, 2008), and people working in the area have come to recognize the need to develop systems with learners in mind and to train systems using learner data.

It should be noted that most ASR research related to learner speech has focused on using the technology for language learning (CALL) rather than on assessment of the type that involves grading. For instance, at the 1998 STiLL workshop in Marholmen, there were just two papers on assessment: Townshend, Bernstein, Todic, and Warren (1998) on the PhonePass system and Cucchiariini, Strik, and Boves (1998) about the Automatic testing of oral proficiency project (Cucchiariini, Strik, & Boves, 2000a, 2000b, 2002). The subsequent InSTiL and SLaTE workshops also paid more attention to learning than to grading.

## Assessment

Assessing speaking fundamentally requires eliciting a language sample and evaluating that language sample. The assessment outcomes can be provided in the form of feedback (in CALL) or in the form of grades (in testing). Where elicitation is concerned, it should be kept in mind that the design of assessment, particularly in the form of test tasks, has always been jointly determined by the construct being measured and by what the constraints of technology and practicality allow. In speaking assessments using ASR-based technology, task types used have generally included reading aloud, elicited imitation, and short free responses. For evaluation, most often, the procedure is as follows: recorded speech is parsed and tagged, and machine scores (predictors, independent variables) for these are obtained (semi-)automatically, and are evaluated by comparing them to one or more reference scores (criteria, dependent variables) obtained manually. The features evaluated relate to repeat accuracy, length of production, fluency (e.g., rate of speech; number and duration of pauses, silences, and disfluencies), vocabulary, grammatical accuracy (by comparing to a reference language model), and pronunciation (by comparing to a reference acoustic model).

Many studies have been conducted to compare the quality of these automated scores with those produced by human markers. In general, the relation between human and automatic grading improves if longer stretches of speech and multiple utterances are used (e.g., Bernstein, 2012; Neumeyer, Franco, Digalakis, & Weintraub, 2000).” For repeat accuracy, Graham, Lonsdale, Kennington, Johnson, and McGhee (2008) used an Elicited Imitation (EI) task, where learners have to listen and then repeat utterances that varied in complexity and length. High correlations (0.92) were observed between automatic and human EI scores. Cook, McGhee, and Lonsdale (2011) used the Oral Proficiency Interview (OPI) to obtain scores on L2 oral proficiency, as have De Wet and colleagues (De Wet, Van Der Walt, & Niesler, 2009; Müller, De

Wet, Van Der Walt, & Niesler, 2009), and found similar results. In all these studies it was found that repeat accuracy can be a good predictor of oral proficiency.

Fluency-related features is the focus of several studies (Cucchiarini *et al.*, 2000b, 2002, 2010; Ginther, Dimova, & Yang, 2010). It was found that automatic measures can be used to predict fluency ratings, but the predictive power of such measures is stronger for read speech than for spontaneous speech. Different speakers and utterances were used, but many similarities were observed: higher for spontaneous speech are articulation rate, mean length of silent pauses, total duration of silent pauses, number of filled pauses; lower for spontaneous speech are rate of speech and phonation/time ratio (obviously, because the number and length of the pauses is larger), and number of broken words.

Some assessment research has also noted the importance of pronunciation in accounting for L2 oral proficiency (e.g., Isaacs, [this issue](#); Plough, Briggs, & Van Bonn, 2010). For this reason, rate of speech has also been compared to other features, such as “goodness of pronunciation” measures, a likelihood ratio that compares the phones produced against the phones expected (Kanters, Cucchiarini, & Strik, 2009; Strik, Truong, De Wet, & Cucchiarini, 2009; Witt & Young, 2000). In Cucchiarini *et al.* (2000a), correlations were calculated between automatic measures and four human scores—an overall score, and three subscores (segmental quality, fluency, and speech rate) for a reading task. In this and in other studies (De Wet *et al.*, 2009; Müller *et al.*, 2009), the correlations for “goodness of pronunciation” measures were generally lower than those for rate of speech and repeat accuracy.

Two of the better known oral proficiency test engines are Versant, which is now used in the Pearson PTE Academic speaking test, and SpeechRater, which the Educational Testing Service (ETS) uses in its TOEFL practice speaking tests. Versant, originally called SET-10 or PhonePass, was originally developed by Ordinate Corporation (Bernstein, 1999; Bernstein & Cheng, 2007; Townshend *et al.*, 1998). The tasks included are reading aloud, repeating sentences, and giving short answers to questions. Machine subscores are calculated on pronunciation, fluency, vocabulary, and sentence mastery, and these subscores are then combined to obtain one overall score. Reported correlations with human raters are 0.84–0.92 for the subscores and 0.92 overall (Bernstein & Cheng, 2007). Similar correlations are reported for Dutch as a target language in De Jong, Lennig, Kerkhoff, and Poelmans (2009). In SpeechRater, various features are extracted from the audio signal and then combined by means of a multiple regression scoring model (Xi, Higgins, Zechner, & Williamson, 2008; Zechner *et al.*, 2014; Zechner, Higgins, Xi, & Williamson, 2009). When all of the learner’s responses are aggregated, a speaker-level correlation of 0.73 is obtained (Zechner *et al.*, 2014).

## Challenges

Some questions that arise from the foregoing include the following: What type of speech should be used in assessments, and how should they be elicited? Which human and machine scores should be considered, and how should such scores be obtained and validated?

Where type of speech is concerned, as previously noted, most studies have used constrained tasks, such as reading aloud and elicited imitation. There have also been some attempts to study less constrained, more spontaneous speech (e.g., Cuccharini *et al.*, 2002; Zechner *et al.*, 2009), but these are of course more difficult for speech technology to handle and the results contain more errors. Whether the domain is constrained (e.g., asking for directions) or open will also affect system performance. Beyond the tasks being more or less constrained, they also need to be at the appropriate level. If the task is too easy, the elicited speech will contain too few errors and there will be little variation in the scores. The task should be at the right level of challenge, so that enough errors are made. For instance, Luo, Minematsu, Yamauchi, and Hiroshi (2009) found that shadowing, a repeat task where test takers only hear the utterance they need to produce, predicts L2 oral proficiency better than tasks where they read the utterance, because it is more demanding. In many cases an adaptive task would be preferable.

Questions also arise regarding the human reference scores and the machine scores. While the number of human scores used in a study is often limited, it is possible to obtain many possible machine scores. Some of these scores can be computed easily by means of standard technology, whereas other scores require dedicated technology that is specifically developed and optimized for these tasks. If the scores can be calculated automatically, and thus can be obtained (several times) for large amounts of speech, testing reliability and validity are less problematic for these scores. The main questions then are which scores should be calculated and how. Because scores can be calculated in different ways, it makes it difficult to make comparisons between studies. What the desired reference measures are will differ from case to case, which can come down to different construct definitions.

Many different types of machine scores were used in the studies previously discussed, and high correlations were obtained in many of them. This seems to imply that various machine scores contribute to the reference score and that there is a large amount of covariation (overlap) between different machine scores. The question then is what to do with these observations. If high correlations are obtained for different machine scores, for different types of speech elicited with different tasks, would it then be sufficient to take one of these procedures (e.g., repeat accuracy in elicited imitation), or should we take a combination of these procedures to obtain a more thorough assessment of different aspects of L2 speaking proficiency? If it's the latter, then the questions are which ones and how to combine them.

Finally, it should be noted that obtaining and interpreting such results (including the correlations) should be done carefully. Details of the procedures used to obtain human and machine scores can influence the results. Therefore, relevant details of these procedures should be clearly described. Even better would be open access to all the data and the software used to calculate the scores, which would improve making comparisons between studies. Because the human scores are used as benchmarks to evaluate the machine scores, these human scores should preferably be evaluated themselves (e.g., by calculating inter- and intrarater agreement and reliability scores) (e.g., see Cucchiaroni *et al.*, 2000b; Stolarova, Wolf, Rinker, & Brielmann, 2014). Furthermore, many of the studies are scientific lab tests based on corpora of recorded speech, and it is well known that results for such controlled lab tests are often (much) better than those obtained in real-life use in practice.

It would appear that the two test providers earlier mentioned have chosen different approaches to dealing with these challenges, by taking different approaches to construct definition and test design and, following that, to determining the suitability of their assessments for various uses. Pearson has chosen to define speaking ability as a psycholinguistic construct (Van Moere, 2012), seeing it as a “real-time activity that requires planning, formulating, articulating, and monitoring” (Downey, Farhady, Present-Thomas, Suzuki, & Van Moere, 2008, p. 162) and that test scores reflect test takers' ability to use “core language component process in real time by quantifying the ease with which the speaker can access and retrieve lexical items, build phrases and clause structures, and articulate responses, without conscious attention to the linguistic code” (Downey *et al.*, 2008, p. 162). Thus defined, the technology does adequately assess the chosen construct and avoids the question and criticism of the mostly constrained tasks not being particularly communicative or interactive (Bernstein, 1999; Chapelle & Chung, 2010; Chun, 2006). The correlation between scores obtained on these tests with scores obtained on oral proficiency interviews is cited as further proof of the technology's readiness.

For its part, ETS defines speaking in communicative terms, as “the use of oral language to interact directly and immediately with others” (Butler, Eignor, Jones, McNamara, & Suomi, 2000, p. 2) in academic settings. The effects of this can be seen in TOEFL speaking test tasks that are more contextualized and less restricted (Pearlman, 2008; Xi *et al.*, 2008), and in a research program that seeks to develop automated measures that account for content, coherence, and interactive speaking (Evanini *et al.*, 2014; Evanini, Xie, & Zechner, 2013; Wang, Evanini, & Zechner, 2013). While more contextualized, the tasks remain monologic rather than interactive, and the less restricted the tasks, the more challenging to evaluate.



Thus, Pearson takes a psycholinguistic approach to construct definition, contends that their ASR-based technology captures that construct of L2 speaking ability sufficiently, and therefore, if one accepts their argument, ready for (and indeed has been deployed in) high-stakes testing; however, ETS takes a more communicative approach to construct definition and, going by that definition, concluded that ASR-based technology is not quite ready for use in high-stakes testing and therefore uses them only in practice tests, subject to further research and improvement.

In any event it is encouraging that the performance of ASR has gradually improved, and recently the use of deep learning through DNNs has led to a boost in performance (e.g., see Pereyra *et al.*, 2016; Xiong *et al.*, 2016), and has also been used for assessing speech (Cheng, Chen, & Metallinou, 2015; Tao, Ghaffarzadegan, Chen, & Zechner, 2016), and their content (Evanini *et al.*, 2013; Malinin, Van Dalen, Wang, Knill, & Gales, 2016; Yoon & Xie, 2014). The possibilities of using ASR-based technology are thus increasing, which could make it possible to shift from relatively simple and constrained tasks to more complex ones, such as the spontaneous speech and dialogic tasks required by communicative constructs, to which we now turn.

### Spoken dialogue systems

In communicatively oriented speaking tests (e.g., IELTS), candidates produce not only individual turns but also engage in question-answering and discussion with an examiner (Seedhouse, Harris, Naeb, & And Ustunel, 2014). It is thus worth seeing to what extent technology—in the form of spoken dialogue systems—can automate candidate-examiner interaction. ASR is the starting point for interactive spoken dialogue systems (SDS), which use both speech and natural language processing technologies to enable extended human-machine conversations. Standard SDS consist of the following set of system components, typically in a pipelined architecture (Figure 2). The ASR component, as described previously, first transcribes a spoken user utterance. Next, the *natural language understanding* component extracts the transcription's syntactic structure and/or meaning, which is used by the *dialogue manager* to determine an appropriate system response. Finally, the *natural language generation* component maps the desired content of the system response to a text string, which is given to the *text-to-speech* component to produce a spoken system

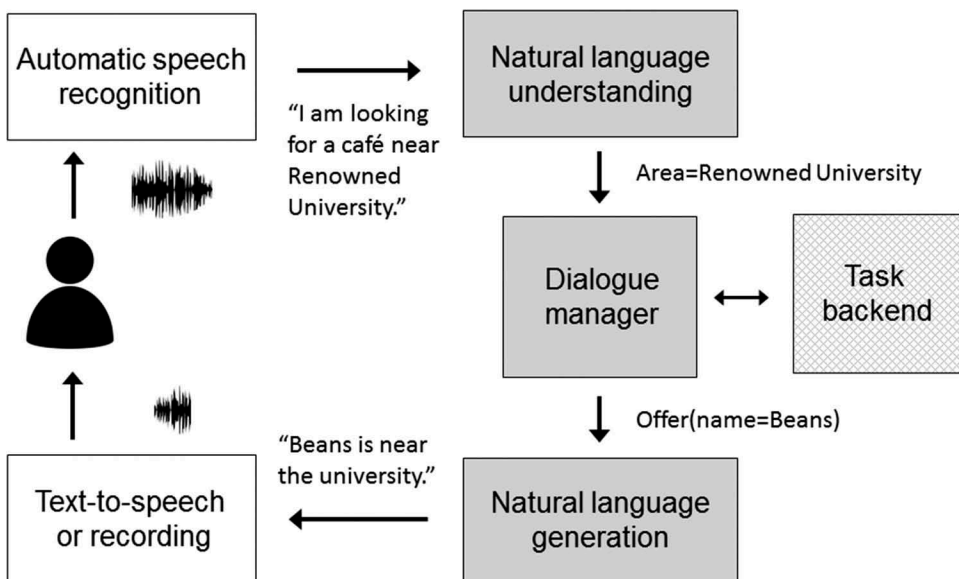


Figure 2. Sample SDS architecture.

utterance. This overview will primarily focus on the dialogue manager of the SDS, which is probably the most challenging component of the system.

Unlike ASR, dialogue management has neither a dominant computational paradigm nor a standard evaluation goal. Methods for building dialogue managers include both hand-crafted and data-driven approaches, while the dialogue managers themselves are formalized by using a wide variety of computational representations. For evaluation, metrics range from subjective assessments, such as usability, to objective criteria, such as rate of task completion or time spent executing a task.

## History of SDS

The earliest conversational systems were text rather than speech based and thus did not include either an ASR or a text-to-speech component. An extremely early but well-known example is the ELIZA program, which simulated a therapist by using simple pattern-matching techniques (Weizenbaum, 1966). As computing power and ASR quality improved (e.g., in supporting real-time recognition of speaker-independent, large vocabulary, continuous speech), researchers started exploring the feasibility of building *spoken* dialogue systems. While the potential benefits of moving from typed to spoken user input include remote or hands-free access, ease of use, and naturalness, spoken input also introduces new technical challenges, such as making a dialogue manager robust to error-prone ASR. As with ASR, much foundational SDS research was funded by DARPA in the 1990s through its Air Travel Information Service (ATIS) program and later through its Communicator program. Since then, commercial SDS have become numerous, with applications ranging from call centers (e.g., AT&T's early "How May I Help You?" system (Gorin, Riccardi, & Wright, 1997)) to current-day intelligent personal assistants, such as Apple's Siri and Microsoft's Cortana for operating systems, and Amazon's Alexa and Google's Assistant for smart speakers.

Although SDS have become common in a variety of well-chosen application areas, language assessment is not yet one of them. Most SDS research still focuses on task-oriented and information-seeking applications, such as providing telephone or microphone access to restaurant (Gasic *et al.*, 2013) or tourist (Singh, Litman, Kearns, & Walker, 2002; Young *et al.*, 2010) information. Nonetheless, there is increasing interest in building SDS for educational applications, ranging from one-on-one tutoring systems for STEM (Forbes-Riley & Litman, 2011) to systems for teaching or assessing the speaking skills of L2 learners in immersion-like situations (Eskensazi, 2009).

As with CALL, recent decades have seen the emergence of several special interest groups, workshops, and conference special sessions focusing on educational applications of speech and language technologies (including SDS). Beginning in the 1990s a series of tutorial dialogue systems workshops began to span the Artificial Intelligence and Education and the Natural Language Processing (NLP) communities, although most work focused on typed rather than spoken dialogue systems and on STEM rather than language domains. In 2003 the first of a series of (North American Chapter of the) Association for Computational Linguistics (NAACL/ACL) workshops on Innovative Use of NLP for Building Educational Applications was held. Besides the already noted formation in 2006 of the ISCA SIG Speech and Language Technology in Education and its associated organization of seven workshops, there have also been related special sessions at Interspeech. However, the proportion of SDS research at these events (both in general and as applied to language learning/assessment in particular) is still rather low. As with CALL, there also seems to be more interest in developing learning rather than assessment SDS technology.

Currently, most automated spoken language training or assessment is based on a learner's response to a stimulus, such as an image or a reading, and is typically noninteractive in that system behavior does not vary on the basis of the learner's prior response(s). Even when using an SDS to perform interactive training or assessment, the language skills being assessed are themselves often noninteractive (e.g., based on utterance-level properties, such as pronunciation, vocabulary, or grammar). For example, in language learning, an interactive SDS was developed to support personalized pronunciation training by using the learner's dialogue history to dynamically pick a next



training sentence that would be optimal for the learner (Su, Wu, & Lee, 2015). Although the system was primarily evaluated with simulated users, two real learners who used the system demonstrated pronunciation improvements from pre- to posttest. Although not a language-learning system per se, Raux and Eskenazi (2004) showed how modifying prompts of a task-oriented SDS based on a user's prior input could implicitly teach non-native speakers how to improve their vocabulary and grammar.

To date, there has been little work in using the interactive capabilities of an SDS to teach and assess interactive language skills themselves. The research in this area has typically involved the construction of scenarios in which learners need to have a successful conversation with a system to achieve a scenario's goals (Litman *et al.*, 2016; Seneff, Wang, & Chao, 2007). While most approaches have involved two-party dialogues, a recent variant aimed at engaging young learners and providing more conversational roles for learners created a triologue-based system where learners interacted with two system partners (Evanini *et al.*, 2014). A preliminary evaluation suggested the promise of using standard SDS components for young language learners and of using triologue tasks to assess conversational competence. Other groups have used SDS technology to help language learners acquire not only language skills but also the cultural skills needed to complete a task via conversational interaction (Johnson & Valente, 2008) and to provide an opportunity for conversational practice via the use of chatbots (Cabral *et al.*, 2014). McGraw and Seneff (2007) suggest that despite the limitations of SDS technology, language learning and assessment applications have properties (e.g., more user tolerance of recognition errors, pedagogical value of misrecognized utterances, usefulness of scenario-guided conversations in narrow domains) that system designers can exploit to yield robust SDS—at least from the speech and language perspective. Finally, although interactive dialogue proficiency is multifaceted, many of these facets are related to current research in SDS (Jurafsky & Martin, 2009). This synergy suggests the promise of using SDS to assess a variety of conversational skills as is discussed below.

## Challenges

Just as the needs of CALL require different approaches to ASR, depending on the speaking construct, significant modification will likely be required of existing approaches to SDS and also to the design of assessments. First, because of technology limitations, conversations with dialogue systems exhibit different characteristics (e.g., they are simpler and more constrained) than conversations with other humans. For example, a timetable dialogue system might open a conversation with “Where are you leaving from?” rather than “How may I help you?” to keep the user's reply within the system's vocabulary and grammar. It is unclear whether dialogue systems will be able to yield the types of user conversational behaviors that one would ideally wish to assess. While dialogue systems technology may be useful for automating routine interactions (e.g., where human examiners currently use standardized scripts to ask questions about familiar topics), automating more open-ended interactive discussions that assessors are interested in will be much more challenging. Thus, the assessor will need to consider the extent to which their construct can accommodate such deviation from authentic dialogues. It should be noted, however, that this challenge exists not just for SDS but also for interactive speaking tests involving human interlocutors, which are also somewhat constrained and routinized (Seedhouse & Egbert, 2006; Seedhouse & Harris, 2011) in service of the requirements of reliability.

Second, in speaking assessment applications, not only the dialogue systems but also the users have limited speaking skills. Language learners are more likely to speak with incorrect pronunciation and to use incorrect lexical and grammatical structures (e.g., “When the bus?” rather than “When is the next bus?”). This will pose challenges for current dialogue system techniques and resources, which are designed to accept speech and language inputs from conversationally proficient users. While the complications for ASR were discussed earlier, similar issues are relevant for all of the other components of a SDS, and dedicated technology will likely again be needed to handle the challenges.

Though, for assessment purposes, a weakness may be turned into strength, and the number of instances and how severely the dialogue breaks down can potentially be a measure of a learner's abilities or lack thereof.

Third, from the perspective of building on tutorial dialogue system technology, another challenge of language learning is that most user utterances—particularly when going beyond routine interactions such as form-filling—will be neither clearly right nor wrong. This will make utterance assessment, and in turn feedback generation, more difficult than tutoring in a well-defined domain. For example, while “friction” is an incorrect physics response to the question “What are the forces exerted on the man after he releases his keys?” and “gravity” is a correct response (Forbes-Riley & Litman, 2011), both are appropriate conversational responses in the sense that they attempt to answer the question and the content is loosely on topic. While current tutorial technology typically builds on task-oriented or information-seeking dialogue systems, systems for language learning might also need to incorporate aspects of chatbot systems (where the goal is not to deeply understand a user but instead to move the conversation forward by providing vague system responses). Fortunately, chatbot research is becoming of increasing interest to the SDS community.

Finally, to be useful in a pedagogical or assessment setting, the SDS will need to be easily configurable by language experts (who are not likely to be proficient in building spoken dialogue systems) to blend into existing structures. While statistical approaches to building SDS can reduce the costs of system configuration and deployment by replacing manual authoring with automatic knowledge acquisition directly from conversational data, a large amount of training data is first needed. Although the SDS community is increasingly making transactions with various task-oriented SDS freely available (e.g., see the download materials at [dialrc.org](http://dialrc.org) or <http://research.microsoft.com/en-us/events/dstc/>), similar corpora suitable for training language-oriented SDS are as yet unavailable. In fact, creating such resources will likely pose challenges for existing corpus collection and annotation methods (e.g., how to change existing SDS systems) to facilitate human scoring (e.g., logging the speech file for the complete dialogue rather than a single side of the conversation) (Litman *et al.*, 2016).

### Opportunities

Despite the challenges noted above, human-human assessment dialogues do share some features with existing computer-human dialogues (e.g., both human assessors and computers use standardized scripts and utterances, as illustrated by the human assessment dialogue excerpt from the IELTS Speaking test in Figure 3). There is also a base of related educational and SDS technology that SDS-based assessment can build on (e.g., automatic assessment systems for spontaneous but nonconversational speech; SDS for tutoring and chatbots). Thus, there are targeted opportunities for performing interactive spoken language assessment in the context of a spoken dialogue system. These include both utterance-level assessment of what a user says and how the user says it (which is already needed to guide the real-time operation of the

**Examiner: Do you work or are you a student**

**Candidate: I'm a student in univeristy er**

**Examiner: And what subject are you studying**

**Figure 3.** Testing dialogue excerpt between an IELTS human examiner and candidate (Seedhouse *et al.*, 2014).

dialogue manager), as well as dialogue-level assessment of conversational properties, such as turn-taking and dialogue structure (which can evaluate a user's conversational abilities by examining multiple utterances).

At the utterance-level, a dialogue manager typically uses the assessments from the speech recognition and natural language understanding components, in conjunction with an internal representation of system state, to decide what the dialogue system should do next. If a finite state machine is used to computationally represent a dialogue manager, the SDS is modeled as being in exactly one of a finite number of states at any given time, with the dialogue manager defined by a list of its states, its initial state, and the conditions for moving from one state to another. A common approach is to have the states correspond to system utterances, with the assessments of user utterances determining the transitions between states.

Figure 4 presents part of an example finite state dialogue manager that could generate the IELTS Speaking test dialogue shown in Figure 3. In Figure 4, the states are shown as rounded boxes, with the initial state highlighted by shadowing. The arrows show how utterance assessment (i.e., the output of the natural language understanding component, recall Figure 2) is used to move from one state to another. Given this particular dialogue manager, a system would begin a dialogue by using natural language generation (recall Figure 2) to generate an utterance associated with the initial state. The examiner's first utterance in Figure 3 is one possibility. Next, the candidate's reply is assessed by the natural language understanding (NLU) component. In this dialogue manager, only a simple meaning assessment is needed to map an utterance to one of two answers desired by the system. For example, the candidate's utterance in Figure 3 would be assessed as job = student. This would cause the system to transition to the Ask (subject) state and generate an associated system utterance; the examiner's second utterance in Figure 3 could be one possibility. Note that speech and natural language processing can be used to assess the speech files and transcriptions representing the user's utterances for many linguistic dimensions. For example, syntactic analysis can be used to detect grammatical errors. If assessment was for syntax rather than meaning, one possible arrow from the start state might have been labeled "NLU: grammatical = NO" and this arrow might have led to a state, such as Ask (repeat-but-grammatically). Semantic analysis can be used to assess meaning for an expected answer at both fine (e.g., paraphrase recognition) and coarse (e.g., on-topic or off-topic recognition) grained levels of analysis. Knowledge of pragmatics can be used to assess skills, such as politeness, whereas knowledge of discourse can be used to evaluate local contextual coherence. Finally, acoustic and prosodic information particular to speech can be used to assess speaking fluency and to improve the recognition of an utterance's conversational function.

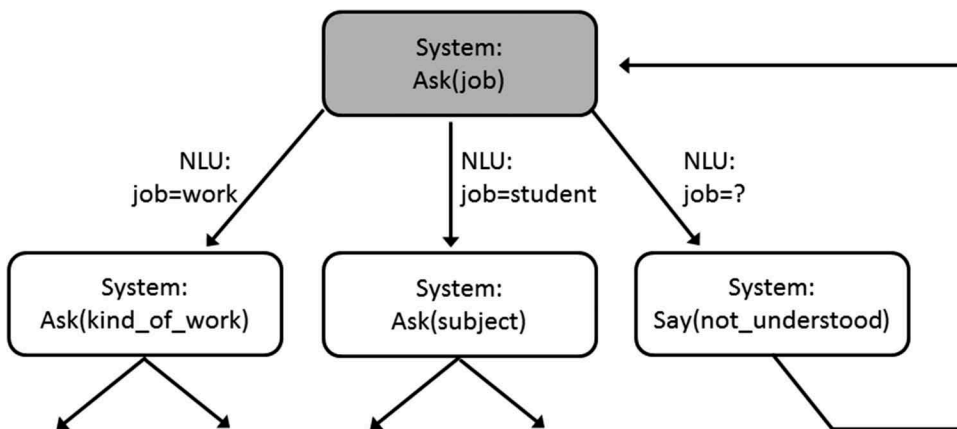


Figure 4. Sample finite state dialogue manager.

While utterance-level assessment in an interactive spoken dialogue system may overlap with existing assessment tasks in noninteractive contexts, there are often differences in moving to a dialogue context. Fortunately, work in both tutorial SDS and in spontaneous speech assessment have provided insights and results to build on. For example, the goal of traditional short answer scoring is to produce a numeric score that agrees with a gold-standard human score, using statistical techniques, such as lexical or semantic similarity, as well as approaches based on deeper semantic processing and inference. In contrast, the goal of short answer assessment in a dialogue system is to use similar methods to assign a label corresponding to an allowable transition from the system's current dialogue state (e.g., in a tutorial dialogue system, assessing a response to a tutor's question as correct, partially correct, or wrong, to reach the dialogue state corresponding to the most appropriate system feedback) (Dzikovska, Nielsen, & Brew, 2012).

Second, and as previously noted for CALL, utterances produced during dialogue are often more spontaneous and unconstrained than utterances produced in noninteractive contexts, making them less predictable and harder to assess on many dimensions. As a result, compared to text, assessment of speech proficiency has focused less on aspects, such as semantics, discourse, and pragmatics where errors in ASR can propagate, and more on aspects, such as pronunciation and fluency. Nonetheless, there have been some promising results on the assessment of spontaneous speech using dimensions based on ASR output (e.g., syntax, discourse) or using holistic scales, such as ones based on the CEFR (e.g., Shashidhar, Pandey, & Aggarwal, 2015; Van Dalen, Knill, & Gales, 2015; Wang *et al.*, 2013; Xiong, Evanini, Zechner, & Chen, 2013; Zechner & Bejar, 2006).

In addition, the interactive capabilities of dialogue systems enable the use of SDS methods, such as dialogue state tracking, to better handle noisy utterance assessments. In contrast to a finite state dialogue manager where the system operates on the basis of being in one state at a time and discards any information regarding the less likely states, with state tracking the dialogue manager instead estimates the probability of transitioning to all possible states. For example, in Figure 4, after processing an utterance, such as "I'm at a university," the system might believe that it is in the student-answer state with 75% probability, the work-answer state with 20% probability, and the nonunderstanding state with 5% probability. This capability provides robustness to ASR errors and to ambiguities that often can be resolved further in the dialogue. State tracking is often implemented by using methods from artificial intelligence such as Bayesian networks (Williams, Raux, Ramachandran, & Black, 2013). Another approach to handling uncertainty is to trigger a system clarification (e.g., Stoyanchev, Liu, & Hirschberg, 2013) when the best assessment of a user's utterance is of low confidence. Finally, although utterance assessments for online dialogue management must be based on linguistic features that can be computed in *real-time*, a wealth of such features exists in the SDS literature.

At the dialogue-level, assessment typically involves higher-level and contextual user abilities that require multiple utterances of the dialogue for analysis and that reflect the fact that dialogue is a joint activity involving two or more conversational participants. For example, in a coherent dialogue, consecutive user utterances should not be isolated and unrelated to one another. Instead, user utterances should exhibit semantic and topical relationships with both the system's and the user's utterance history (e.g., Gandhe & Traum, 2008). In addition, user utterances should be used to achieve appropriate conversational functions, such as providing an answer after a system question, or ending the dialogue with a closing rather than a greeting (e.g., Cuayáhuil, Dethlefs, Hastie, & Lemon, 2013). Users should also be able to use linguistic devices, such as referring expressions, connectives, prosody, etc., that are both consistent with the underlying relationships between utterances and that are used at appropriate times during the conversation (e.g., Gravano, Hirschberg, & Beňuš, 2012). For turn-taking abilities, users should be able to both recognize when it is their turn in a dialogue and use linguistic signals to convey to the system that they are maintaining or ending their turn (e.g., Raux & Eskenazi, 2009). Users must also be able to effectively and collaboratively coordinate the conversation with the system (e.g., Visser, Traum, DeVault, & Op Den Akker, 2014), making it clear what the user has actually heard and understood, generating

confirmations to the system when necessary, and appropriately recovering from system misunderstandings.

Most research in the area of SDS has focused on understanding the human dialogue abilities that were enumerated above to build better spoken dialogue systems, rather than to assess user behavior along conversational dimensions. Nonetheless, there are approaches being developed to evaluate the quality of simulated (i.e., computer) users of a spoken dialogue system which work in SDS that language assessment could build upon. Prior work, for example, has evaluated simulated users for features, such as quantity of user activity, distribution of dialogue functions of user utterances, and overall success and efficiency of the interaction (e.g., Ai & Litman, 2008). Evaluation frameworks have similarly been developed to evaluate the quality of dialogue systems (e.g., for optimizing user satisfaction) (Walker, Litman, Kamm, & Abella, 1997). Such evaluation approaches could potentially be adapted to assess the dialogue abilities of human partners from their interactions with spoken dialogue systems.

For example, in routine scenarios, whether and how quickly the dialogue resulted in task completion could be assessed, while for any conversation the quality of the dialogue could be assessed by examining user response times to questions, utterance content for topic and conversational function, presence of overlapping speech, number of clarification and repetition requests by both conversational partners, and so on.

## Conclusion

Significant progress has been made in ASR and SDS so that these systems are now beginning to be used for assessing language learners' speaking abilities, albeit the language produced and assessed in these assessments, whether monologic or dialogic, are on the whole relatively constrained.

To what extent such assessments adequately assess speaking ability depends in part on how the construct is defined. Some have defined speaking ability in psycholinguistic terms and see that the features measured by present-day systems do capture the construct. Others define speaking ability in communicative terms as involving interactional competence, and for these there is some way to go before such assessments can be used in higher-stakes contexts. Ultimately, the validity of tests for particular uses depends on appropriate validity arguments being made and evidence being shown for them (Kane, 2006). There are no perfect tests, because test design is always about balancing different desired qualities, not all of which can be jointly maximized (Saville, 2003). Like technology-based speaking assessments have their limitations, so also do human-mediated speaking tests. Thus, the validity of the former, as with the latter, will depend on all the other elements that make up the assessment. Indeed, it should be noted that the above presume constructs that see speaking as human-to-human interaction. However, increasingly, people are also having spoken interactions with computers, and the day may come when what we want to assess is the ability to communicate both with humans and with machines, in which case the validity equation for using these technologies changes completely.

But before that day, the further application of these technologies to speaking assessment would be facilitated by more communication and collaboration between the language assessment and spoken language technologies communities. Particularly useful would be the development of conversational scenarios that respect the constraints and best practices of both communities, the collection of associated corpora of learner-SDS interactions, and the manual scoring of such interactions to produce "gold-standard" assessments that are the target for automation. For example, a recent proof-of-concept study (Litman *et al.*, 2016) demonstrated the feasibility of (a) using existing SDS to collect dialogues with non-native speakers of English, (b) human-assessing CEFR levels in such SDS speech, and (c) using an automated assessment system designed for prompted but noninteractive speech to yield assessment scores that can positively correlate with humans. More research along these lines would not only facilitate the formulation of more well-defined interdisciplinary research tasks but would generate important technical resources for the data-driven design and evaluation of automated systems for spoken language assessment that more fully capture the construct for the future.



## ORCID

Gad S. Lim  <http://orcid.org/0000-0001-5208-4953>

## References

- Ai, H., & Litman, D. J. (2008). Assessing dialog system user simulation evaluation measures using human judges. *Proceedings of 46th Annual Conference of Association of Computational Linguistics*, 622–629. Columbus, Ohio, USA: Association for Computational Linguistics.
- Bernstein, J. (1999). *PhonePass testing: Structure and construct*. Menlo Park, CA: Ordinate.
- Bernstein, J. (2012). Computer scoring of spoken responses. In C. A. Chapelle (Ed.), *Encyclopedia of applied linguistics* (pp. 857–863). New York, NY, USA: Wiley.
- Bernstein, J., & Cheng, J. (2007). Logic, operation and validation of a spoken english Test,” chapter 8. In V. M. Holland & F. P. Fisher (Eds.), *Speech technologies for language learning* (pp. 174–194). New York, NY, USA: Routledge.
- Butler, F. A., Eignor, D., Jones, S., McNamara, T., & Suomi, B. K. (2000). *TOEFL 2000 speaking framework: A working paper*. TOEFL Monograph Series MS-20. Princeton, NJ: Educational Testing Service.
- Cabral, C., Campbell, N., Ganesh, S., Gilmartin, E., Haider, F., Kenny, E., ... Orosko, O. R. (2014). *MILLA – A multimodal interactive language agent*. Edinburgh, United Kingdom: eNTERFACE, Software.
- Chapelle, C. A., & Chung, Y. R. (2010). The promise of NLP and speech processing technologies in language assessment. *Language Testing*, 27(3), 301–315. doi:10.1177/0265532210364405
- Cheng, J., Chen, X., & Metallinou, A. (2015). Deep neural network acoustic models for spoken assessment applications. *Speech Communication*, 73, 14–27. doi:10.1016/j.specom.2015.07.006
- Chun, C. (2006). Commentary: An analysis of a language test for employment: The authenticity of the phonepass test. *Language Assessment Quarterly*, 3(3), 295–306. doi:10.1207/s15434311laq0303\_4
- Coniam, D. (1999). Voice recognition software accuracy with second language speakers of English. *System*, 27(1), 49–64. doi:10.1016/S0346-251X(98)00049-9
- Cook, K., McGhee, J., & Lonsdale, D. (2011). Elicited imitation for prediction of OPI test scores. *Proceedings of the Sixth Workshop on Innovative Use of NLP for Building Educational Applications*, 30–37. Portland, Oregon, USA: Association for Computational Linguistics.
- Cuayahuitl, H., Dethlefs, N., Hastie, H., & Lemon, O. (2013). Impact of ASR N-Best information on bayesian dialogue act recognition. *Proceedings of SIGDIAL*. Metz, France: Association for Computational Linguistics.
- Cucchiarini, C., Strik, H., & Boves, L. (1998). Automatic pronunciation grading for Dutch. *Proceedings of the ESCA Workshop on Speech Technology in Language Learning*, 95–98. Marholmen, Sweden: ESCA.
- Cucchiarini, C., Strik, H., & Boves, L. (2000a). Different aspects of expert pronunciation quality ratings and their relation to scores produced by speech recognition algorithms. *Speech Communication*, 30, 109–119. doi:10.1016/S0167-6393(99)00040-0
- Cucchiarini, C., Strik, H., & Boves, L. (2000b). Quantitative assessment of second language learners’ fluency by means of automatic speech recognition technology. *Journal of the Acoustical Society of America*, 107(2), 989–999. doi:10.1121/1.428279
- Cucchiarini, C., Strik, H., & Boves, L. (2002). Quantitative assessment of second language learners’ fluency: Comparisons between read and spontaneous speech. *Journal of the Acoustical Society of America*, 111(6), 2862–2873. doi:10.1121/1.1471894
- Cucchiarini, C., Van Doremalen, J., & Strik, H. (2010). Fluency in non-native read and spontaneous speech. *Proceedings of the DiSS-LPSS Joint Workshop 2010*, 15–18. Tokyo, Japan: University of Tokyo.
- De Jong, J. H. A. L., Lennig, M., Kerkhoff, A., & Poelmans, P. (2009). Development of a test of spoken Dutch for prospective immigrants. *Language Assessment Quarterly*, 6(1), 41–60. doi:10.1080/15434300802606564
- De Jong, N. (this issue). Fluency in second language testing: Insights from different disciplines. *Language Assessment Quarterly*.
- De Wet, F., Van Der Walt, C., & Niesler, T. R. (2009). Automatic assessment of oral language proficiency and listening comprehension. *Speech Communication*, 51, 864–874. doi:10.1016/j.specom.2009.03.002
- Delcloque, P. (Ed.) (1999). Progressing interface transparency: Speech applications in computer assisted language learning. *Proceedings of ‘Integrating Speech Technology in Learning’ (InSTIL)*, Besançon, France.
- Delcloque, P. (Ed.) (2000). Speech technology in language learning and the assistive interface. *Proceedings of ‘Integrating Speech Technology in Learning’ (InSTIL)*, Dundee, Scotland.
- Delmonte, R. (2004) InSTIL/ICALL Symposium ‘NLP and Speech Technologies in Advanced Language Learning Systems’, Venice, Italy. <http://project.cgm.unive.it/events/ICALL2004/index.htm>
- Derwing, T. M., Munro, M. J., & Carbonaro, M. (2000). Does popular speech recognition software work with ESL speech? *TESOL Quarterly*, 34(3), 592–603. doi:10.2307/3587748
- Downey, R., Farhady, H., Present-Thomas, R., Suzuki, M., & Van Moere, A. (2008). Evaluation of the usefulness of the Versant for English test: A response. *Language Assessment Quarterly*, 5(2), 160–167. doi:10.1080/15434300801934744



- Dzikovska, M. O., Nielsen, R. D., & Brew, C. (2012). Towards effective tutorial feedback for explanation questions: A dataset and baselines. *Proceedings Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT '12)*, 200–210. Montreal, Canada: Association for Computational Linguistics.
- Eskenazi, M. (2009). An overview of spoken language technology for education. *Speech Communication*, 51, 10. doi:10.1016/j.specom.2009.04.005
- Evanini, K., So, Y., Tao, J., Zapata-Rivera, D., Luce, C., Battistini, L., & Wang, X. (2014). Performance of a trialogue-based prototype system for English language assessment for young learners. *Proceedings of the Interspeech Workshop on Child Computer Interaction (WOCCI)*. Singapore, Singapore: ISCA.
- Evanini, K., Xie, S., & Zechner, K. (2013). Prompt-based content scoring for automated spoken language assessment. *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*, 157–162. Atlanta, Georgia, USA: Association for Computational Linguistics.
- Forbes-Riley, K., & Litman, D. (2011). Benefits and challenges of real-time uncertainty detection and adaptation in a spoken dialogue computer tutor. *Speech Communication*, 53(9–10), 1115–1136. doi:10.1016/j.specom.2011.02.006
- Galaczi, E., & Taylor, L. (this issue). Interactional competence: Conceptualisations, operationalisations, and outstanding questions. *Language Assessment Quarterly*.
- Gandhe, S., & Traum, D. (2008). An evaluation understudy for dialogue coherence models. *Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue*, 172–181. Columbus, Ohio, USA: Association for Computational Linguistics.
- Gasic, M., Breslin, C., Henderson, M., Kim, D., Szummer, M., Thomson, B., ... Young, S. (2013). POMDP-based dialogue manager adaptation to extended domains. *Proceedings of SIGDIAL*, 214–222. Metz, France: Association for Computational Linguistics.
- 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. doi:10.1177/0265532210364407
- Gorin, A. L., Riccardi, G., & Wright, J. H. (1997). How may I help you? *Speech Communication*, 23, 113–127. doi:10.1016/S0167-6393(97)00040-X
- Graham, C. R., Lonsdale, D., Kennington, C., Johnson, A., & McGhee, J. (2008). Elicited imitation as an oral proficiency measure with ASR scoring. *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008)*, 1604–1610. Marrakech, Morocco: LREC.
- Gravano, A., Hirschberg, J., & Beňuš, Š. (2012). Affirmative cue words in task-oriented dialogue. *Computational Linguistics*, 38(1), 1–39. doi:10.1162/COLI\_a\_00083
- Isaacs, T. (this issue). Shifting sands in second language pronunciation assessment research and practice. *Language Assessment Quarterly*.
- Johnson, W. L., & Valente, A. (2008). Tactical language and culture training systems: Using artificial intelligence to teach foreign languages and cultures. *Association for the Advancement of Artificial Intelligence*, 30(2), 1632–1639.
- Jurafsky, D., & Martin, J. (2009). *Speech and language processing: An introduction to natural language processing, speech recognition, and computational linguistics* (2nd ed.). New Jersey, USA: Prentice-Hall.
- Kane, M. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. pp. 17–64). Westport, CT: Praeger.
- Kanters, S., Cucchiari, C., & Strik, H. (2009). The goodness of pronunciation algorithm: A detailed performance study. *Proceedings of the 2009 ISCA Workshop on Speech and Language Technology in Education (SLaTE)*, 2–5. Austin, Texas, USA: ISCA.
- Litman, D., Young, S., Gales, M., Knill, K., Ottewell, K., Van Dalen, R., & Vandyke, D. (2016). Towards using conversations with spoken dialogue systems in the automated assessment of non-native speakers of english. *17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 270–275. Los Angeles, California, USA: SIGdial.
- Luo, D., Minematsu, N., Yamauchi, Y., & Hiroshi, K. (2009). Analysis and comparison of automatic language proficiency assessment between shadowed sentences and read sentences. In *Proceedings of SLaTE* (pp. 2009). Warwick, England: ISCA.
- Malinin, A., Van Dalen, R. C., Wang, Y., Knill, K. M., & Gales, M. J. F. (2016). Off-topic response detection for spontaneous spoken english assessment. *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL 2016)*, 1075–1084. Berlin, Germany: Association for Computational Linguistics.
- McGraw, I., & Seneff, S. (2007). Immersive second language acquisition in narrow domains: A prototype ISLAND dialogue system. *Proceedings of SLaTE*, 84–87. Farmington, Pennsylvania, USA: ISCA.
- Müller, P., De Wet, F., Van Der Walt, C., & Niesler, T. (2009). Automatically assessing the oral proficiency of proficient L2 speakers. *Proceedings of SLaTE*, 29–32. Austin, Texas, USA: ISCA.
- Neumeyer, L., Franco, H., Digalakis, V., & Weintraub, M. (2000). Automatic scoring of pronunciation quality. *Speech Communication*, 30, 83–94. doi:10.1016/S0167-6393(99)00046-1
- Pearlman, M. (2008). Finalizing the test blueprint. In C. A. Chapelle, M. K. Enright, & J. M. Jamieson (Eds.), *Building a validity argument for the test of English as a foreign language* (pp. 227–258). New York, NY, USA: Routledge.

- Pereyra, G., Tucker, G., Chorowski, J., Kaiser, L., & Hinton, G. (2016). Regularizing neural networks by penalizing confident output distributions. arXiv:1701.06548v1 [cs.NE]; <https://arxiv.org/abs/1701.06548>
- Plough, I., Briggs, S., & Van Bonn, S. (2010). A multi-method analysis of evaluation criteria used to assess the speaking proficiency of graduate student instructors. *Language Testing*, 27(2), 235–260. doi:10.1177/0265532209349469
- Raux, A., & Eskenazi, M. (2004). Non-Native Users in the Let's Go!! Spoken Dialogue System: dealing with linguistic mismatch. *Proceedings of HLT-NAACL*, 217–224. Boston, Massachusetts, USA: Association for Computational Linguistics.
- Raux, A., & Eskenazi, M. (2009). A finite-state turn-taking model for spoken dialog systems. *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 629–637. Boulder, Colorado, USA: Association for Computational Linguistics.
- Saville, N. (2003). The process of test development and revision within UCLES EFL. In C. J. Weir & M. Milanovic (Eds.), *Continuity and innovation: revising the cambridge proficiency in english examination 1913-2002* (pp. 57–120). Cambridge, United Kingdom: UCLES/Cambridge University Press.
- Seedhouse, P., & Egbert, M. (2006). The interactional organization of the IELTS speaking test. *IELTS Research Reports*, 6, 161–204.
- Seedhouse, P., & Harris, A. (2011). Topic development in the IELTS speaking test. *IELTS Research Reports*, 12, 69–124.
- Seedhouse, P., Harris, A., Naeb, R., & And Ustunel, E. (2014). The relationship between speaking features and band descriptors: A mixed methods study. *IELTS Research Reports Online Series*, 2, 1–30.
- Seneff, S., Wang, C., & Chao, C. Y. (2007). Spoken dialogue systems for language learning. *Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, 13–14. Rochester, New York, USA: Association for Computational Linguistics.
- Shashidhar, V., Pandey, N., & Aggarwal, V. (2015). Automatic spontaneous speech grading: A novel feature derivation technique using the crowd. *Proceedings 53rd Annual Meeting of the Association for Computational Linguistics and 7th International Joint Conference on Natural Language Processing*, 1085–1094. Beijing, China: Association for Computational Linguistics.
- Singh, S., Litman, D., Kearns, M., & Walker, M. (2002). Optimizing dialogue management with reinforcement learning: Experiments with the NJFun system. *Journal of Artificial Intelligence Research*, 16, 105–133.
- Stolarova, M., Wolf, C., Rinker, T., & Briellmann, A. (2014). How to assess and compare inter-rater reliability, agreement and correlation of ratings: An exemplary analysis of mother-father and parent-teacher expressive vocabulary rating pairs. *Frontiers in Psychology*, 5, 509. doi:10.3389/fpsyg.2014.00509
- Stoyanchev, S., Liu, A., & Hirschberg, J. (2013). Modelling human clarification strategies. *Proceedings of SIGDIAL*, 137–141. Metz, France: SIGdial.
- Strik, H., Neri, A., & Cucchiari, C. (2008). Speech technology for language tutoring. *Proceedings of LangTech 2008*, 73–76. Rome, Italy: LangTech.
- Strik, H., Truong, K., De Wet, F., & Cucchiari, C. (2009). Comparing different approaches for automatic pronunciation error detection. *Speech Communication*, 51, 845–852. doi:10.1016/j.specom.2009.05.007
- Su, P. H., Wu, C. H., & Lee, L. S. (2015). Recursive dialogue game for personalized computer-aided pronunciation training. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(1), 127–141.
- Tao, J., Ghaffarzadegan, S., Chen, L., & Zechner, K. (2016). Exploring deep learning architectures for automatically grading non-native spontaneous speech. *Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2016)*. Shanghai, China: IEEE.
- Townshend, B., Bernstein, J., Todici, O., & Warren, E. (1998). Estimation of spoken language proficiency. *Proceedings of the ESCA Workshop STiLL: 'Speech Technology in Language Learning'*, 179–182. Marholmen, Sweden: ESCA.
- Van Compernelle, D. (2001). Recognizing speech of goats, wolves, sheep and non-natives. *Speech Communication*, 35, 71–79. doi:10.1016/S0167-6393(00)00096-0
- Van Dalen, R., Knill, K., & Gales, M. (2015). Automatically grading learners' English using a Gaussian process. *Proceedings Sixth Workshop on Speech and Language Technology in Education (SLaTE)*, 7–12. Leipzig, Germany: ISCA.
- Van Doremalen, J., Cucchiari, C., & Strik, H. (2010). Optimizing automatic speech recognition for low-proficient non-native speakers. *EURASIP Journal on Audio, Speech, and Music Processing*, 2010(2010), Article ID 973954, 13 pages. doi:10.1186/1687-4722-2010-973954
- Van Doremalen, J., Cucchiari, C., & Strik, H. (2011). Speech technology in CALL: The essential role of adaptation. *Interdisciplinary Approaches to Adaptive Learning: Communications in Computer and Information Science Series*, 26, 56–69.
- Van Moere, A. (2012). A psycholinguistic approach to oral language assessment. *Language Testing*, 29(3), 325–344. doi:10.1177/0265532211424478
- Visser, T., Traum, D., DeVault, D., & Op Den Akker, R. (2014). A model for incremental grounding in spoken dialogue systems. *Journal on Multimodal User Interfaces*, 8(1), 61–73.
- Walker, M. A., Litman, D. J., Kamm, C. A., & Abella, A. (1997). PARADISE: A framework for evaluating spoken dialogue agents. *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*, 271–280. Madrid, Spain: Association for Computational Linguistics.

- Wang, X., Evanini, K., & Zechner, K. (2013). Coherence modeling for the automated assessment of spontaneous spoken responses. *Proceedings Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 814–819. Atlanta, Georgia, USA: Association for Computational Linguistics.
- Weizenbaum, J. (1966). ELIZA—A computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1), 36–45. doi:10.1145/365153.365168
- Williams, J., Raux, A., Ramachandran, D., & Black, A. (2013). The dialog state tracking challenge. *Proceedings of the SIGDIAL 2013 Conference*, 404–413. Metz, France: SIGdial.
- Witt, S. M., & Young, S. (2000). Phone-level pronunciation scoring and assessment for interactive language learning. *Speech Communication*, 30, 95–108. doi:10.1016/S0167-6393(99)00044-8
- Xi, X., Higgins, D., Zechner, K., & Williamson, D. M. (2008). Automated scoring of spontaneous speech using SpeechRater v1.0. Educational Testing Service Research Report No. RR-08-62. Princeton, NJ: ETS.
- Xiong, W., Droppo, J., Huang, X., Seide, F., Seltzer, M., Stolcke, A., ... Zweig, G. (2016). Achieving human parity in conversational speech recognition. arXiv:1610.05256 [cs.CL]; <https://arxiv.org/abs/1610.05256>
- Xiong, W., Evanini, K., Zechner, K., & Chen, L. (2013). Automated content scoring of spoken responses containing multiple parts with factual information. *Proceedings SLATE 2013*, 137–142. Porto, Portugal: ISCA.
- Yoon, S., & Xie, S. (2014). Similarity based non-scorable response detection for automated speech scoring. *Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*, 116–123. Baltimore, Maryland, USA: Association for Computational Linguistics.
- Young, S., Gasic, M., Keizer, S., Mairesse, F., Schatzmann, J., Thomson, B., & Yu, K. (2010). The hidden information state model: A practical framework for POMDP-based spoken dialogue management. *Computer Speech and Language*, 24, 150–174. doi:10.1016/j.csl.2009.04.001
- Zechner, K., & Bejar, I. (2006). Towards automatic scoring of non-native spontaneous speech. *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, 216–223. New York, NY, USA: Association for Computational Linguistics.
- Zechner, K., Evanini, K., Yoon, S. Y., Davis, L., Wang, X., Chen, L., ... Leong, C. W. (2014). Automated scoring of speaking items in an assessment for teachers of English as a Foreign Language. *Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications*, 134–142. Baltimore, Maryland, USA: Association for Computational Linguistics.
- 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. doi:10.1016/j.specom.2009.04.009