Language Learning & Technology
ISSN 1094-3501
February 2021, Volume 25, Issue 1
pp. 44–47

BOOK REVIEW

Review of *Automated speaking assessment: Using language technologies to score spontaneous speech*

Yasin Karatay, *Iowa State University*

Leyla Karatay, *Iowa State University*

Automated speaking assessment: Using language technologies to score spontaneous speech

Klaus Zechner & Keelan Evanini (Eds.)

2020

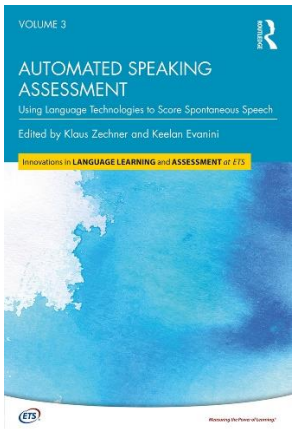
ISBN: 978-1-13805-687-9

US $ 43.96

211 pp.

Routledge

New York, NY



In this dynamic technology world, where innovative ideas and algorithms evolve almost daily, the landscape of computer-based language testing over the years has been dramatically reformed (Suvorov & Hegelheimer, 2013). One unique challenge that remains, however, is to effectively assess speaking proficiency (Chapelle & Voss, 2016), and since reliability plays a fundamental role in test score interpretation, the degree of error in speaking scores should be taken into account by test developers (Chapelle, 2021, p. 72). In oral speaking assessment, human rating, as traditionally the only means to evaluate constructed responses, can pose a threat to validity because of factors such as rater fatigue and rater bias (Zechner & Evanini, 2020, p. 3). A possible approach to addressing these potential threats is automated speech recognition (ASR). Recent developments such as the use of Deep Neural Network algorithms for training ASR and growth in the performance of scoring models have considerably advanced automated speech processing technology. Yet, there is still a lot to be discovered because of the lack of knowledge about how these systems work. Responding to this need, *Automated Speaking Assessment: Using Language Technologies to Score Spontaneous Speech*, edited by Zechner and Evanini, aims to bridge the gap between stakeholders such as scientists and engineers who may not have any expertise in reliability, validity, and fairness issues and stakeholders in academia who may not have expertise in speech processing technology. In doing so, the book unveils the state-of-the-art automated speech scoring technology, SpeechRater, used at Educational Testing Service (ETS). This book aims to be easily accessible to graduate students and applied linguists as well as testing professionals and measurement specialists.

Structurally, the book is comprised of 11 chapters organized into four main parts based on the relevance of topical content. The first three chapters in "Part I: Introduction" serve as an overview of the key aspects of automated scoring before diving into a broader psychometric context of the main technical components of an automated speech scoring system introduced in Chapters 4 to 5 in "Part II: Components of an Automated Speech Scoring System." The three chapters (Chapters 6 to 8) in "Part III: Speech Features" present a more detailed overview of the feature selection process from the perspective of Applied Linguistics and Second Language Acquisition. Finally, the remaining three chapters (Chapters 9 to 11) in "Part IV: Recent Developments and Outlook" discuss future developments in automated scoring technology.

As the book editors, Zechner and Evanini start the first chapter by explaining its necessity. They also present the areas where automated scoring systems are used, followed by a brief overview of the systems at ETS.

In order to help readers understand the automated scoring systems in the subsequent chapters, the authors then define key concepts. Chapter 2 presents a wide range of published validation frameworks and highlights the importance of contextualized validation. In this chapter, Zhang, Bridgeman, and Davis suggest that while using automated speech scoring, its context of use should be part of validity evidence. They state that the context of use includes, but is not limited to, the nature of assessment and how the scores will be used. Another central issue that this chapter elucidates is using human ratings as validation criteria in the process of development and assessment of automated speech scoring systems, drawing the readers' attention to the limitations of human ratings. In Chapter 3, Zhang, Yao, Haberman, and Dorans provide a thorough discussion on evaluating the reliability of the scoring model built in the SpeechRater system. In doing so, they present an exhaustive description of a detailed probabilistic foundation by providing an empirical analysis of human scores, SpeechRater scores, and a combination of both and demonstrates the relative contributions of each approach.

Part II introduces the main technical components of an automated speech scoring system. Chapter 4 pertains to ASR as the first component of automated speech scoring systems. After a short presentation of the history of ASR systems, Qian, Lange, and Evanini introduce the concept of ASR in a manner that targets a non-technical audience. The main components described in this chapter are speech signal representation, the acoustic model, the pronunciation dictionary, the language model, and the decoder. The authors also underline the difficulties that can be faced while using ASR, such as audio quality, the number of speakers the system is trained on, and the problems related to non-native speech. Chapter 5 presents two approaches for building scoring models: machine learning models and filtering models. In the first approach, Loukina and Yoon indicate that a single score is assigned to a performance based on different algorithms, such as decision trees, multiple regression, and deep learning. In the latter approach, the authors explain that test takers' responses that have atypical features are automatically detected and flagged not to be scored using the scoring model. The takeaway message in this chapter is that progress in score model performance is non-linear, meaning that building more complex machine learning algorithms may not necessarily yield more reliable results.

The focus of Part III is on the SpeechRater features that represent the construct coverage of the TOEFL iBT speaking test: delivery, language use, and topic development. In Chapter 6, Hsieh, Zechner, and Xi thoroughly discuss how assessing fluency and pronunciation through SpeechRater yields a robust representation of the construct because they do not depend on the textual ASR. In Chapter 7, Yoon, Lu, and Zechner point out that measuring the language use aspect of spontaneous speech through automated syntactic analysis like part of speech (POS) tagging and sentence parsing is typically more difficult than the delivery features for the ASR system. This is because of the frequent disfluencies and grammatical errors, especially in non-native speech. The authors also discuss some ways to evaluate spoken language by adapting the approaches specifically designed for evaluating written language use features. Finally, Chapter 8 focuses on the topic development dimension of the speaking construct and how it is unsatisfactorily operationalized in automated speech scoring systems. Here Wang and Evanini emphasize that the current approaches for processing the semantic and pragmatic aspects of human language fall short in producing reliable measurements of the target features. Complementing the discussion about the language use features in Chapter 7, the authors argue that discourse-related features also heavily rely on the preciseness of the ASR output. It is suggested that the nature of non-native spontaneous speech and the complexity of identifying accurate clause and sentence boundaries add to the problem.

The last part of the book starts with Chapter 9, which contains a detailed explanation of how to provide feedback to learners using speech scoring systems in TOEFL Practice Online. Based on a case study, Gu and Davis describe some of the features of the SpeechRater they selected to use for feedback and the decision-making process of creating feedback reports. They subsequently present future recommendations to investigate the validity of feedback based on the features of SpeechRater. In Chapter 10, Ramanarayanan, Evanini, and Tsuprun introduce spoken dialogue systems (SDS) and help the readers look into automated scoring systems from another perspective. The authors argue that SDSs offer a more detailed assessment of speaking ability by providing interactive, conversational speaking tasks. In an attempt to introduce the

technical components of SDSs, this chapter presents a case study on a free, online TOEFL test preparation course, which uses SDS tasks designed for language learning purposes. It then provides insight into the possible ways in which automated scoring systems, such as SpeechRater, can be used to evaluate test-takers' interactional competence on SDS-based dialogic tasks. The final chapter written by Zechner provides a summary of the book and the automated speech scoring research. Since automated scoring systems are new and bear some challenges, the author proposes several suggestions for future researchers to overcome those challenges and discusses the hybrid rating model (i.e., combined computer-human rating) as a potential solution for broad construct coverage.

Overall, by delineating the complexities and challenges in the conceptualization and operationalization of the automated speech scoring systems, this edited volume manages to lift the veil on automated assessment, improving the validity of machine scoring. While doing so, the contributors to this timely book, who comprise research scientists, programmers, statisticians, applied linguists, and experts in language assessment at ETS, candidly share their insiders' views, with detailed descriptions and implications, as well as rich critical reflections. As the first book that takes an all-inclusive approach to speech technologies for automated scoring of constructed responses in assessing oral speaking, this edited volume is an excellent contribution to the field of language testing, specifically computer-assisted language testing. Readers of this book will substantially benefit from an in-depth under-the-hood description of the basic technologies behind automated speech scoring, feedback, and SDSs and likely appreciate the discussion on technical quality as well as challenges, validity, and fairness in language assessment. Another strength of the book is that the authors do not shy away from the fact that construct coverage of speaking can be limited by today's technology and enhanced by the developments in the foreseeable future. Finally, the book enriches existing research on SDSs, which the authors see as the pioneering technology to elicit spoken responses from language learners in interactive speaking tasks, thus potentially enabling a complete assessment of speaking ability.

Despite the book's positive attributes, some limitations must be mentioned. First, the sheer breadth of content covered, as well as the depth, pertinent to the probabilistic models in Chapter 3, can be overwhelmingly complicated for some readers. While the book was intended to be an essential reference for researchers, developers, and graduate students in artificial intelligence, educational technology, and language learning and assessment as described in the series editors' foreword, the inconsistent level of technicality throughout the chapters may undermine the readability for some of the target audience. Another problem that contributes to the issue of readability is the insufficient use of figures, especially when describing complex systems. For example, similar to Figure 10.1 (p. 178) that illustrates the schematic of the technical components of the HALEF SDS, another figure would help readers visualize the blueprint of an ASR system on page 62. Last but not least, as the main premise of the book is to provide insights into spontaneous speech scoring, one would expect the discussion of assessing spontaneous conversational dialogic speech to be presented earlier rather than at the end of the book. In addition to the location of this important aspect of the speaking construct, the amount of the discussion of issues related to interactional competence such as turn-taking strategies, the appropriate use of different registers, aspects of pragmatic competence may imply that only a part of the construct-like delivery and language use on monologic tasks has been emphasized for the sake of promoting the effectiveness of SpeechRater. It should be noted, however, that the editors and authors of various chapters acknowledge that one reason why monologic speech was the focus of the associated topic is because of assessment practicality and logistics. They also emphasize that the long-term goal of spoken language assessment should be to assess spontaneous conversational dialogic speech as well. However, even though the authors do not shy away from technology's failure to cover a broad construct of speaking, the way this issue is handled in the book may underestimate the complete coverage of the speaking construct.

Despite these weaknesses, this volume is currently the most comprehensive introduction to the use of automated scoring of spontaneous oral speaking performance of language learners and test takers. The book accomplishes what it aimed for by serving as a reference book for graduate students, applied linguists, and measurement specialists interested in automated speaking assessment.

References

Chapelle, C. A. (2021). *Argument-based validation in testing and assessment*. Sage Publishing.

Chapelle, C. A., & Voss, E. (2016). 20 years of technology and language assessment in Language Learning & Technology. *Language Learning & Technology, 20*(2), 116–128. http://dx.doi.org/10125/44464

Suvorov, R., & Hegelheimer, V. (2013). Computer‐assisted language testing. In A. J. Kunnan (Ed.), *The Companion to Language Assessment* (pp. 594–613). John Wiley & Sons, Inc. https://doi.org/10.1002/9781118411360.wbcla083

About the Authors

Yasin Karatay is a Ph.D. student and a research assistant in the Applied Linguistics and Technology program at Iowa State University. His research interests include computer-based speaking assessment, CALL use in materials development and assessment, and English for Specific Purposes.

E-mail: ykaratay@iastate.edu

Leyla Karatay is a Ph.D. student in the Applied Linguistics and Technology program at Iowa State University. She has taught a variety of tertiary-level courses at Duzce University in Turkey. She is interested in computer-assisted language testing, specifically the assessment of integrated listening and speaking skills.

E-mail: lkaratay@iastate.edu