THE JOURNAL OF ASIA TEFL
Vol. 17, No. 3, Fall 2020, 1156-1157
http://dx.doi.org/10.18823/asiatefl.2020.17.3.34.1156

The Journal of Asia TEFL

http://journal.asiatefl.org/

e-ISSN 2466-1511

Book Review

Automated Speaking Assessment: Using Language Technologies to Score Spontaneous Speech, by Klaus Zechner and Keelan Evanini (Eds.), New York, Routledge, 2020, 212 pp., 36.89£ (Ebk), ISBN 978-1-315-16510-3

Automated Speaking Assessment: Using Language Technologies to Score Spontaneous Speech is edited by two experts in automated speech scoring at ETS, Klaus Zechner and Keelan Evanini. It is the third volume in the series entitled *Innovations in Language Learning and Assessment at ETS*. Reflecting two decades of ETS research on automated speaking assessment, the volume provides an insightful overview and update on this increasingly important scientific domain.

The chapters are organized into four sections to guide the reader through the theoretical and technical issues involved in the development and evaluation of speech technologies for automated scoring.

In Part 1 the two editors provide an overall picture of the field of automated speech scoring, followed by two chapters addressing two central concepts of language testing: validity and reliability. Zhang, Bridgeman, and Davis in Chapter 2 focus on a number of key ideas relating to validating automated speech scoring systems after reviewing the existing theoretical validation frameworks. Instead of building a new validation model, they highlight critical validity issues like contextualized validation, human ratings as validation criteria, and the treatment of unusual responses. Chapter 3 addresses reliability through an empirical study of the scoring and assessment accuracy of three different scoring approaches using human scores, SpeechRater scores, and a combination of both. Within this study the authors introduce a number of new psychometric concepts, including "true score" as a better evaluation criterion than a single observed human score, the Proportional Reduction of Mean Squared Error (PRMSE) metric for evaluating reliability, and the Best Linear Predictor approach for designing statistical scoring models.

In order to build valid and reliable automated scoring systems, it is crucial to obtain accurate Automatic Speech Recognition (ASR) and design effective Filtering and Scoring Models. Part II is devoted to the introduction of these indispensable components. Chapter 4 focuses on Automatic Speech Recognition (ASR), providing information about the technical components of ASR systems, describing factors affecting their accuracy, and discussing the potential uses of the output of an ASR system. Two models that operate after feature extraction, i.e., a filtering model and a scoring model, are interpreted in Chapter 5. The chapter provides a detailed descriptive account of the design process for SpeechRater's scoring model, as well as other approaches to scoring models. In addition, it discusses how to identify and score non-scorable responses in one filtering model, which will have a strong impact on the validity of automated scores.

Part III is concerned with the core of automated speech scoring—feature extraction. In particular the authors describe the SpeechRater features, which are organized according to the three construct dimensions of the TOEFL iBT Speaking scoring rubric: Delivery (fluency, pronunciation); Language Use (vocabulary and grammar); and Topic Development (content, discourse coherence). These three chapters first explain the three construct dimensions and then delineate the features that can be extracted to represent different aspects of them. Chapter 6 reviews features that are related to Delivery, namely fluency, segmental pronunciation and suprasegmental pronunciation. These features are regarded as the easiest to extract and have the longest history in the development of automated speech scoring.

Comparatively speaking, Language Use and Topic Development, which are discussed in Chapters 7 and 8 respectively, are more difficult to assess reliably, and the chapter authors draw on the experience of automated writing evaluation.

Part IV introduces new directions for automated speech scoring. Previous chapters have mainly focused on SpeechRater, the scoring system developed by ETS to score spontaneous monologue in a testing and assessment context. Chapter 9 shifts the focus to the language learning context, and investigates how to design proper feedback reports to suit learners' needs. Many online applications make use of automated speech scoring for language learning and reading tutoring and assessment, but this popular commercial development is not supported by solid academic research. In Chapter 9 Gu and Davis address a series of issues related to designing feedback reports, including how to select specific features for feedback, how to design feedback reports, and how teachers and learners perceive feedback. Chapter 10 represents an early endeavor in exploring automated scoring of spoken dialogue. Ramanarayanan, Evanini, and Tsuprun delineate the main technical components of a Spoken Dialogue System (SDS), and illustrate how SDS-based speaking tasks can be designed for language learning purposes. This is a preliminary overview of the development steps for a set of SDS tasks in an online TOEFL test preparation course. In the last chapter, the editors of the book discuss the challenges and future development of automated speech scoring.

This book is the first comprehensive monograph on automated speaking assessment. As an operational manual for automated speech scoring, it offers an accessible, under-the-hood description of the basic technologies. Researchers can draw on their experience to facilitate the future development of automated speaking assessments. A significant benefit of this book is its simple but clear language, which attracts not only specialists in the field but also readers who may not have background knowledge of automated scoring.

As a volume edited by ETS researchers and developers, the book mainly approaches automated speaking assessment from the test-designer perspective. Nearly all the chapters focus on the test development stage, but do not address utilization or the consequences for other groups of stakeholders such as test-takers or teachers. In addition, the studies and discussions in this book are largely based on research using ETS automated speech scoring software, i.e. SpeechRater. As the title indicates, this book features a distinct focus on language technologies for open-ended, spontaneous speech. Automated speaking assessment varies greatly in terms of assessment tasks and scoring methods. It remains open to question whether the findings derived from SpeechRater can be generalized to other assessment contexts.

Overall, this volume specifies the technologies of nearly all the components of automated speaking assessments for the first time, and grounds these issues within a theoretical discussion of language assessment and psychometrics. The collaborations between language testing researchers, computational linguists, speech and natural language processing scientists, and psychometricians make this book into a unique contribution to the development of automated scoring.

Manman Gao
School of Foreign Studies, Anhui University, China
Email: gaomm03@foxmail.com

(Received May 18, 2020; Reviewed August 12, 2020; Accepted September 10, 2020)