Automated assessment of fluency and pronunciation in spontaneous speech: Implications for automated speech scoring

Ching-Ni Hsieh & Klaus Zechner

Educational Testing Service

2019 EALTA Conference

Dublin, Ireland

7/31/2019

Background

1. Automated systems are widely used for evaluating highly predictable or constrained speech (e.g., read aloud, sentence repeat) in language assessment.
2. Automated scoring of spontaneous speech (i.e., open-ended, less predictable speech) has been sparse, because of the difficulty in:
   - Automatically transcribing L2 speech
   - Scoring unpredictable speech
3. This talk focuses on the evaluation of fluency and pronunciation features for an automated system for scoring spontaneous speech.

Evaluating automated scoring systems
(Williamson et al., 2012)

1. **Construct relevance and representation**
   - Match between intended construct and automated scoring capability
   - Match between automated generated features and the scoring criteria
2. **Empirical performance**
   - Agreement between automated scores and human scores
     - Automated scoring systems are often modeled to predict human ratings
     - Human ratings are typically used as an evaluation criterion

Strengths and weaknesses of human scoring (Zhang, 2013)

1. **Strengths of human raters**
   - Can cognitively process the information given in a response
   - Can understand and judge the quality of the content
   - Can evaluate discourse coherence and organization
2. **Human-rater errors and biases**
   - May vary in severity or leniency
   - May understand or interpret scoring rubrics inconsistently
   - May apply scoring criteria inconsistently over time (rater drift)
   - May make mistakes due to cognitive limitations

Strengths and weaknesses of automated scoring (Zhang, 2013)

1. **Strengths of automated scoring**
   - Can consistently apply the same scoring criteria across responses and over time
   - Can achieve greater objectivity than human raters
2. **Weaknesses of automated scoring**
   - Can generally evaluate a relatively narrow range of spoken skills
   - Cannot directly evaluate content accuracy or relevance and discourse organization

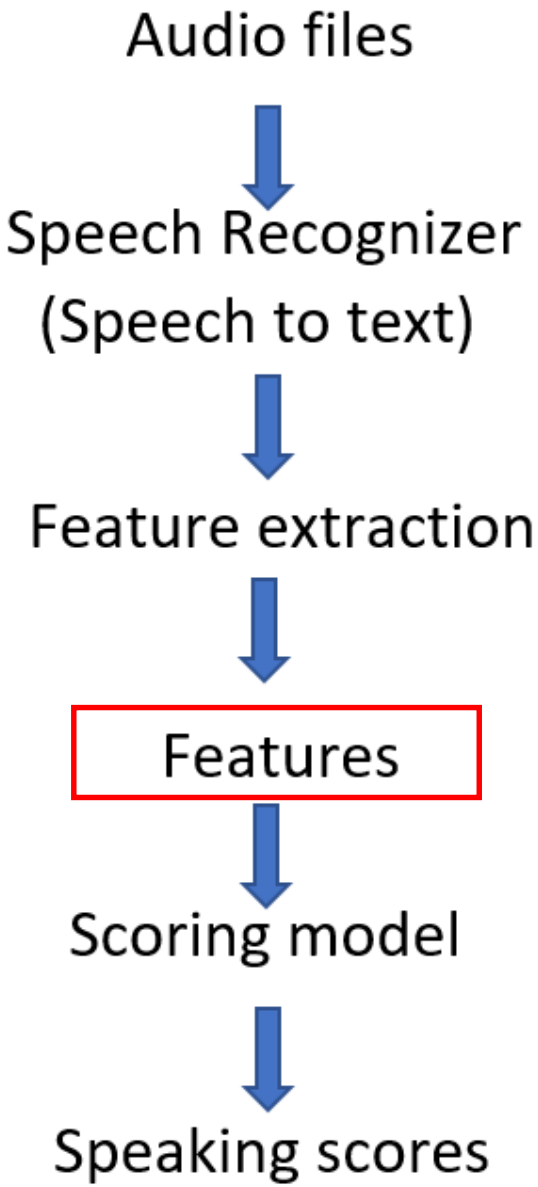
Purposes of the study

- To evaluate construct coverage of fluency and pronunciation features of spontaneous speech generated by *SpeechRater*
- To determine how well *SpeechRater* fluency and pronunciation features correlate with holistic scores of TOEFL iBT Speaking responses awarded by human raters

SpeechRater

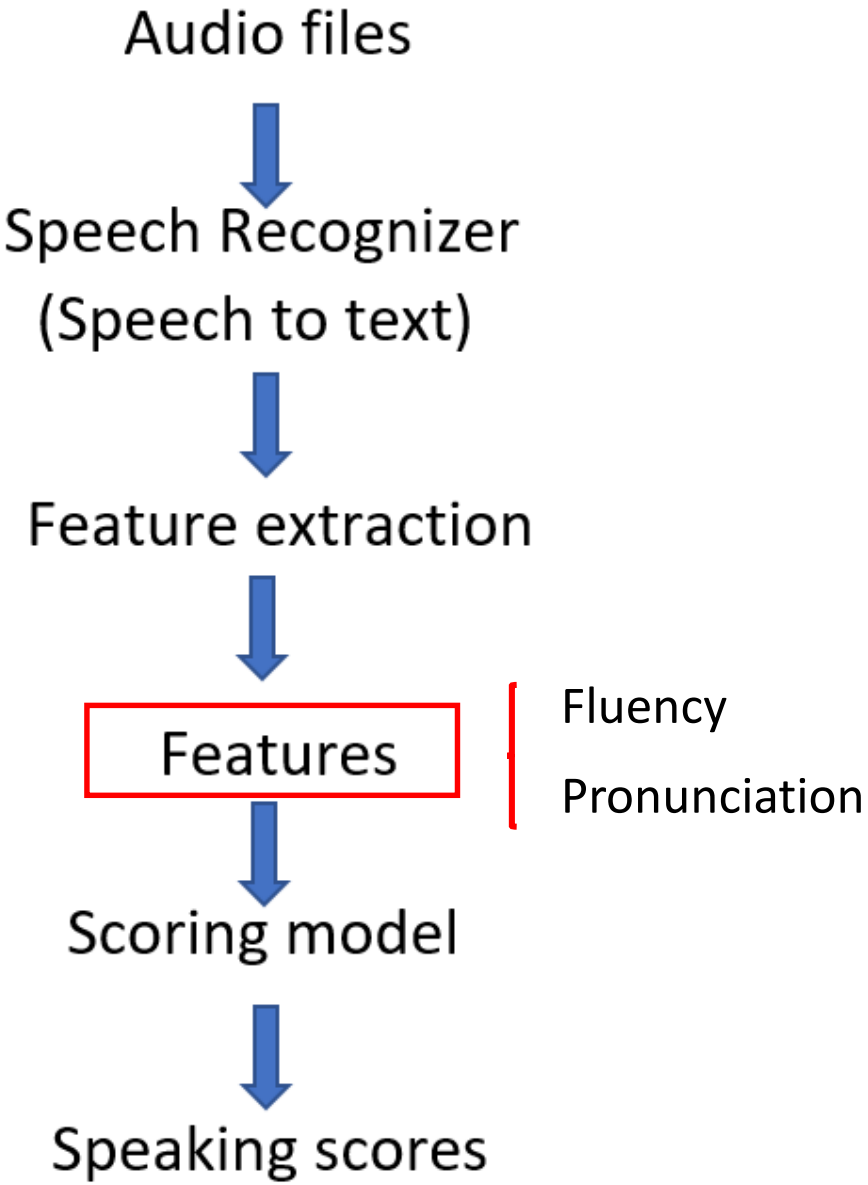
- Automated scoring engine targeted to score open-ended, less predictable L2 speech
- Developed based on a broad conception of speaking proficiency: **Fluency**, **pronunciation**, grammar, vocabulary, topic development
- Has been used to score responses of TOEFL Practice Online (TPO) practice tests since 2006

SpeechRater



***N* = 170+**

SpeechRater



Data set

Data set information	*N*
TOEFL iBT test takers	**38,107**
TOEFL iBT spoken responses (6/test taker)	**228,642**
First languages	121
Native countries	183
Male speakers	18,978
Female speakers	18,319

TOEFL iBT speaking rubrics

SCORE	GENERAL DESCRIPTION	DELIVERY	LANGUAGE USE	TOPIC DEVELOPMENT
4	The response fulfills the demands of the task, with at most minor lapses in completeness. It is highly intelligible and exhibits sustained, coherent discourse. A response at this level is characterized by all of the following:	Generally well-paced flow (fluid expression). Speech is clear. It may include minor lapses, or minor difficulties with pronunciation or intonation patterns, which do not affect overall intelligibility.	The response demonstrates effective use of grammar and vocabulary. It exhibits a fairly high degree of automaticity with good control of basic and complex structures (as appropriate). Some minor (or systematic) errors are noticeable but do not obscure meaning.	Response is sustained and sufficient to the task. It is generally well developed and coherent; relationships between ideas are clear (or clear progression of ideas).
3	The response addresses the task appropriately but may fall short of being fully developed. It is generally intelligible and coherent, with some fluidity of expression, though it exhibits some noticeable lapses in the expression of ideas. A response at this level is characterized by at least two of the following:	Speech is generally clear, with some fluidity of expression, though minor difficulties with pronunciation, intonation, or pacing are noticeable and may require listener effort at times (though overall intelligibility is not significantly affected).	The response demonstrates fairly automatic and effective use of grammar and vocabulary, and fairly coherent expression of relevant ideas. Response may exhibit some imprecise or inaccurate use of vocabulary or grammatical structures or be somewhat limited in the range of structures used. This may affect overall fluency, but it does not seriously interfere with the communication of the message.	Response is mostly coherent and sustained and conveys relevant ideas/information. Overall development is some-what limited, usually lacks elaboration or specificity. Relationships between ideas may at times not be immediately clear.
2	The response addresses the task, but development of the topic is limited. It contains intelligible speech, although problems with delivery and/ or overall coherence occur; meaning may be obscured in places. A response at this level is characterized by at least two of the following:	Speech is basically intelligible, though listener effort is needed because of unclear articulation, awkward intonation, or choppy rhythm/pace; meaning may be obscured in places.	The response demonstrates limited range and control of grammar and vocabulary. These limitations often prevent full expression of ideas. For the most part, only basic sentence structures are used successfully and spoken with fluidity. Structures and vocabulary may express mainly simple (short) and/or general propositions, with simple or unclear connections made among them (serial listing, conjunction, juxtaposition).	The response is connected to the task, though the number of ideas presented or the development of ideas is limited. Mostly basic ideas are expressed with limited elaboration (details and support). At times relevant substance may be vaguely expressed or repetitious. Connections of ideas may be unclear.
1	The response is very limited in content and/ or coherence or is only minimally connected to the task, or speech is largely unintelligible. A response at this level is characterized by at least two of the following:	Consistent pronunciation, stress and intonation difficulties cause considerable listener effort; delivery is choppy, fragmented, or telegraphic; frequent pauses and hesitations.	Range and control of grammar and vocabulary severely limit or prevent expression of ideas and connections among ideas. Some low-level responses may rely heavily on practiced or formulaic expressions.	Limited relevant content is expressed. The response generally lacks substance beyond expression of very basic ideas. Speaker may be unable to sustain speech to complete the task and may rely heavily on repetition of the prompt.

Fluency construct relevance and representation

Construct	Linguistic Phenomenon	SpeechRater feature
Breakdown fluency	Filled pause rate	# filled pauses (uh, um) per second
	Pause duration	Mean duration of pauses in seconds
	Pause frequency	# pauses/total #words

Construct	Linguistic Phenomenon	SpeechRater feature
Speed fluency	Speaking rate	#words per second in total response time
	Articulation rate	#words per second in total articulation time
	Length of run	Mean length of run in words

Construct	Linguistic Phenomenon	SpeechRater feature
Repair fluency	Repetition rate	#repetitions/total # words
	Repair rate	Repair interruption points/ total #words

Fluency construct relevance and representation

Construct	Linguistic phenomenon	SpeechRater feature
Breakdown fluency	Filled pause rate	# filled pauses (uh, um) per second
	Pause duration	Mean duration of pauses in seconds
	Pause frequency	# pauses/total #words

Construct	Linguistic phenomenon	SpeechRater feature
Speed fluency	Speaking rate	#words per second in total response time
	Articulation rate	#words per second in total articulation time
	Length of run	Mean length of run in words

Construct	Linguistic phenomenon	SpeechRater feature
Repair fluency	Repetition rate	#repetitions/total # words
	Repair rate	Repair interruption points/ total #words

Pronunciation construct relevance and representation

Construct	Linguistic Phenomenon	SpeechRater feature
Segmental pronunciation	Global pronunciation	Acoustic model score
	Vowel duration	Differences in vowel durations

Construct	Linguistic Phenomenon	SpeechRater feature
Suprasegmental pronunciation	Stress frequency	Frequency of stressed syllables
	Stress distance	Distances between stressed syllables
	Syllable duration (Rhythm)	Variability of syllable durations

Pronunciation construct relevance and representation

Construct	Linguistic phenomenon	SpeechRater feature
Segmental pronunciation	Global pronunciation	Acoustic model score
	Vowel duration	Differences in vowel durations

Construct	Linguistic phenomenon	SpeechRater feature
Suprasegmental pronunciation	Stress frequency	Frequency of stressed syllables
	Stress distance	Distances between stressed syllables
	Syllable duration (Rhythm)	Variability of syllable durations

Empirical performance: Fluency feature correlations with human scores

Construct	SpeechRater feature	Correlation
Breakdown fluency	Filled pause rate	-.23
	Pause duration	-.32
	Pause frequency	-.50

Construct	SpeechRater feature	Correlation
Speed fluency	Speaking rate	.54
	Articulation rate	.38
	Length of run	.45

Construct	SpeechRater feature	Correlation
Repair fluency	Repetition rate	-.28
	Repair rate	-.26

Empirical performance: Pronunciation feature correlations with human scores

Construct	SpeechRater feature	Correlation
Segmental pronunciation	Global pronunciation	.39
	Vowel duration	-.40

Construct	SpeechRater feature	Correlation
Suprasegmental pronunciation	Stress frequency	.38
	Stress distance	-.47
	Syllable duration (rhythm)	-.40

Discussion: Fluency

1. Some breakdown fluency and speed fluency features have moderately high correlations with human scores.
2. Correlations for filled pauses and repair fluency are lower.
   - Filled pauses and repairs are hard to correctly identify; even for L1 speech

Discussion: Pronunciation

1. Some stress and rhythm features have moderately high correlations with human scores.
2. Features of intonation, not shown here, have relatively low correlations.
   - Identifying the presence or absence of tone events is very challenging for L2 spontaneous speech

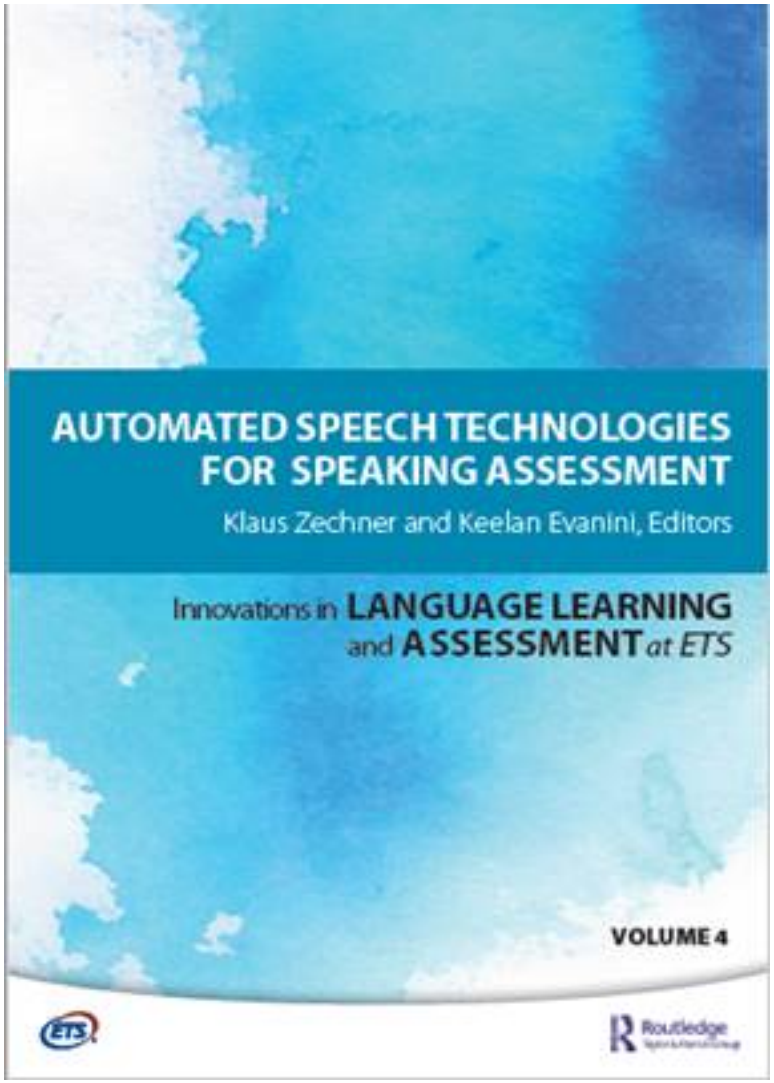
Future research and development

1. Develop more accurate detection of filled pauses and their distributions and repair fluency features.
2. Develop features that measure appropriateness of intonation contours.
3. Improve the accuracy of Automated Speech Recognition (ASR) or reduce the ASR error rate (current data error rate: 20% ; goal: to reduce error rate to 10-15%)

Implications

1. Automated scoring systems can report *holistic* scores and/or *analytic* scores on different speaking constructs, e.g., fluency score, pronunciation score.
2. Automated scoring tools can be used in classrooms to support teaching and learning and help maximize the time and resources available for instruction.

Hsieh, C.-N., Zechner, K., & Xi, X. (in press) Delivery (Fluency and pronunciation). In K. Zechner & K. Evanini (Eds.) *Automated speech technologies for speaking assessment.* New York: Routledge.



Thank you!