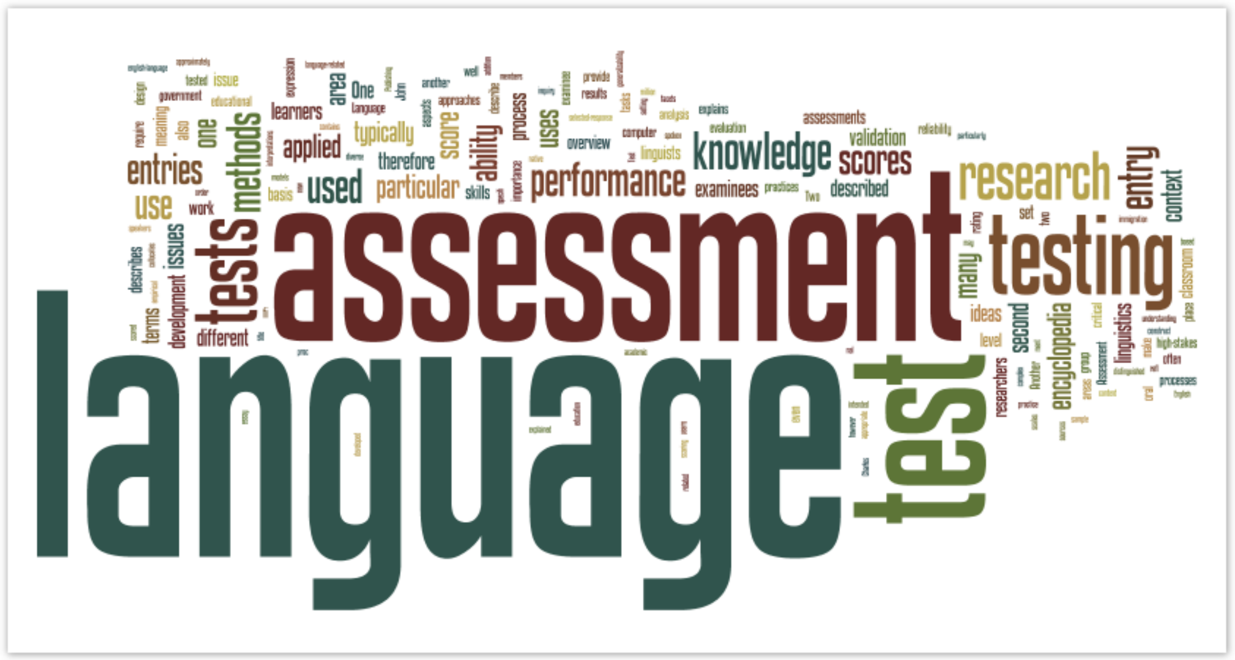
Spoken Language Technology for Language Learning & Assessment



Vikram Ramanarayanan, Klaus Zechner & Keelan Evanini

Educational Testing Service R&D

(*joint work with multiple collaborators at ETS San Francisco and Princeton*)

AGENDA

PART 1: Acoustic, speech, NLP, linguistics basics. (20 min) - VIKRAM

PART 2: State of the art (60 minutes) – KEELAN & KLAUS

BREAK

PART 3: Bleeding Edge (25 minutes) - VIKRAM

PART 4: Q&A + Live Interactive Session/Demos (60 minutes) - ALL

WHAT THIS WORKSHOP *IS* ABOUT…

- Automated scoring basics
- An industry perspective on how to get started with designing automated scoring systems…
- …that can be deployed on actual tests of spoken English…
- …with an outlook on developing technologies.

WHAT THIS WORKSHOP *IS* *NOT* ABOUT…

- Language learning theory
- Assessment theory
- Psychometrics & Validity
- Linguistic theories of discourse
- Machine learning and statistics
- Elaborate software coding

WHO ARE WE AND WHAT DO WE DO?

Group of researchers, engineers, psychometricians, data analysts in San Francisco, CA and Princeton, NJ.

We design, deploy and analyze spoken language technologies for learning and assessment.

Benefits of Spoken Language Technology for Language Learning & Assessment

- Language assessment applications
  - cost reductions (especially for large-scale assessments)
  - faster score reporting
  - increased score consistency and reliability
  - sub-scores / feedback about specific aspects of speaking ability

- Language learning applications
  - ability to practice speaking when no instructor is available
  - targeted, personalized feedback for individual learners
  - interactive, authentic speaking tasks using SDS

How is Spoken Language Technology Currently Being Used for LLA?

- Language assessment applications
  - automated speech scoring system used as sole score
    - Pearson Test of English – Academic, Duolingo English Test, TrueNorth Speaking Assessment (Emmersion), etc.
  - automated speech scoring system combined with human ratings
    - TOEFL iBT Speaking (SpeechRater), Linguaskill (Cambridge English)

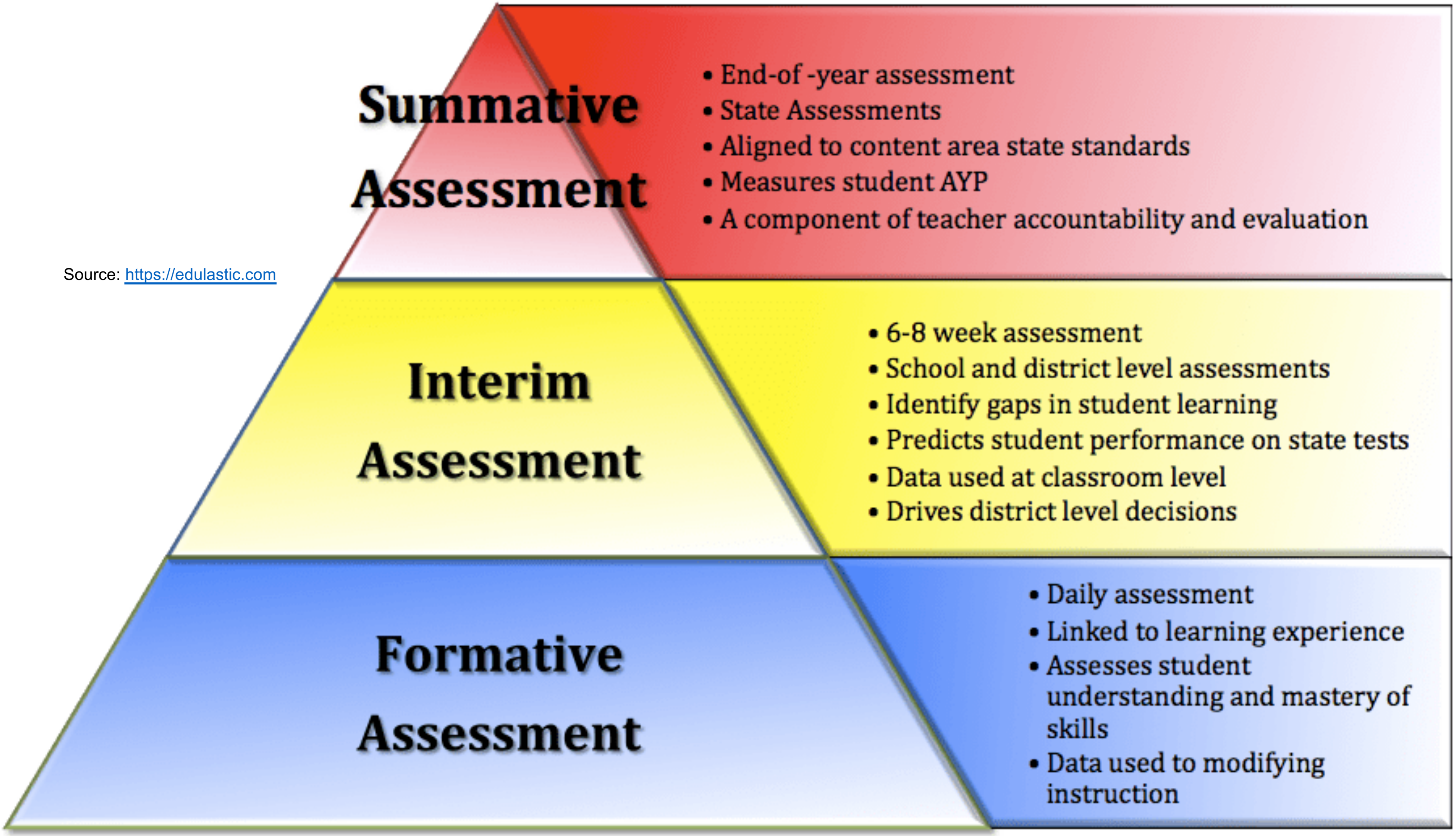
- Language learning applications
  - pronunciation feedback: Carnegie Speech, ELSA, etc.
  - interactive conversations: Alelo, Supiki, etc.

WHAT TERMS DO NEWCOMERS TO L2 SPEECH ASSESSMENT NEED TO KNOW ABOUT?

Vikram Ramanarayanan

10/26/20

FORMATIVE vs SUMMATIVE ASSESSMENT



Summative Assessment

- End-of -year assessment
- State Assessments
- Aligned to content area state standards
- Measures student AYP
- A component of teacher accountability and evaluation

Interim Assessment

- 6-8 week assessment
- School and district level assessments
- Identify gaps in student learning
- Predicts student performance on state tests
- Data used at classroom level
- Drives district level decisions

Formative Assessment

- Daily assessment
- Linked to learning experience
- Assesses student understanding and mastery of skills
- Data used to modifying instruction

Source: https://edulastic.com

PSYCHOMETRICS

- Objective measurement of skills and knowledge, abilities, attitudes, personality traits, and educational achievement.

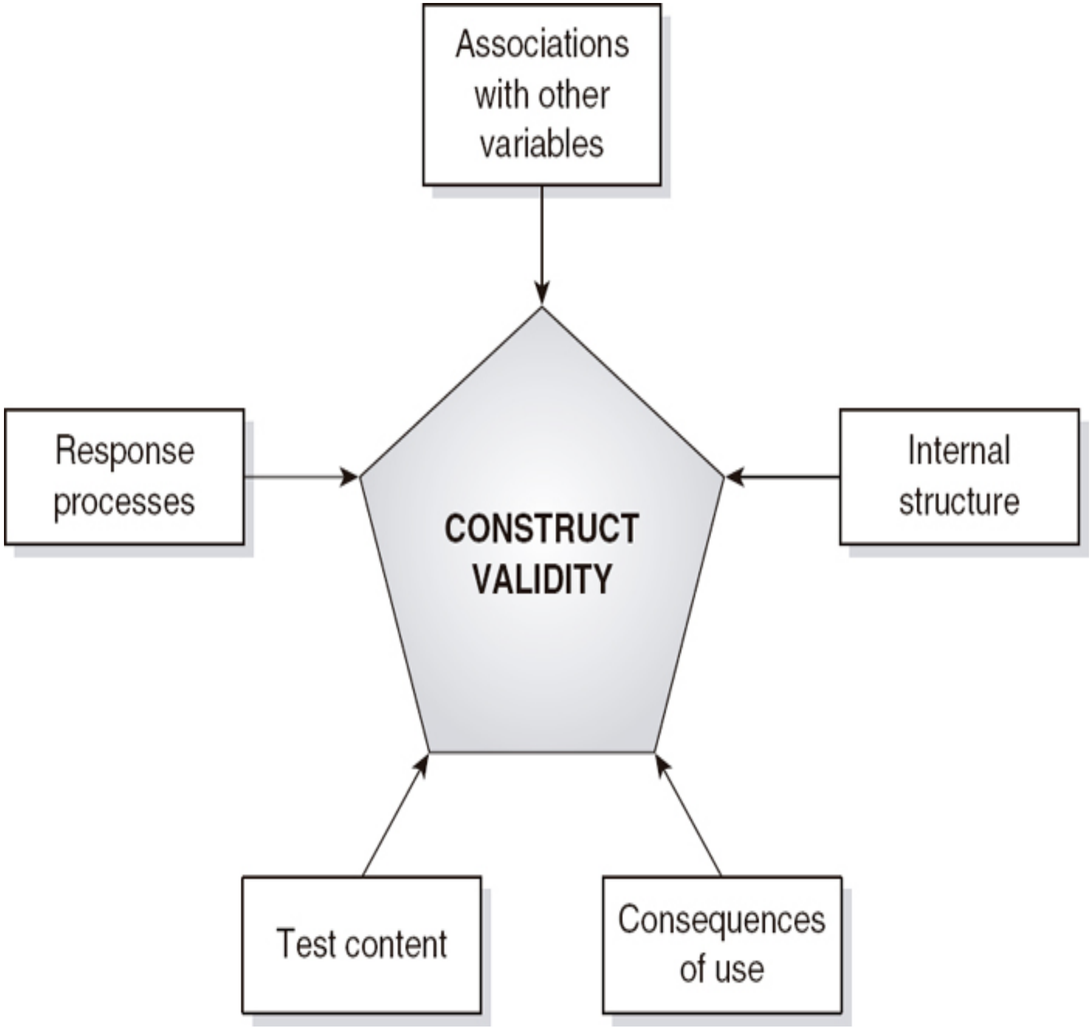
- Key concepts in test theory:
  - Reliability: A reliable measure is one that measures a construct consistently across time, individuals, and situations.
  - Validity: A valid measure is one that measures what it is intended to measure.
  - Reliability is necessary, but not sufficient, for validity.
  - Fairness: do automated scores show any bias towards sub-groups, e.g., gender, native language etc.
  - All can be assessed statistically, and contribute to the quality of an assessment

CONSTRUCT

- **Construct** is the set of knowledge, skills and abilities that a given assessment is designed to provide information about

- Should NOT depend on the scoring approach

- Should depend on the language use domain



RUBRIC

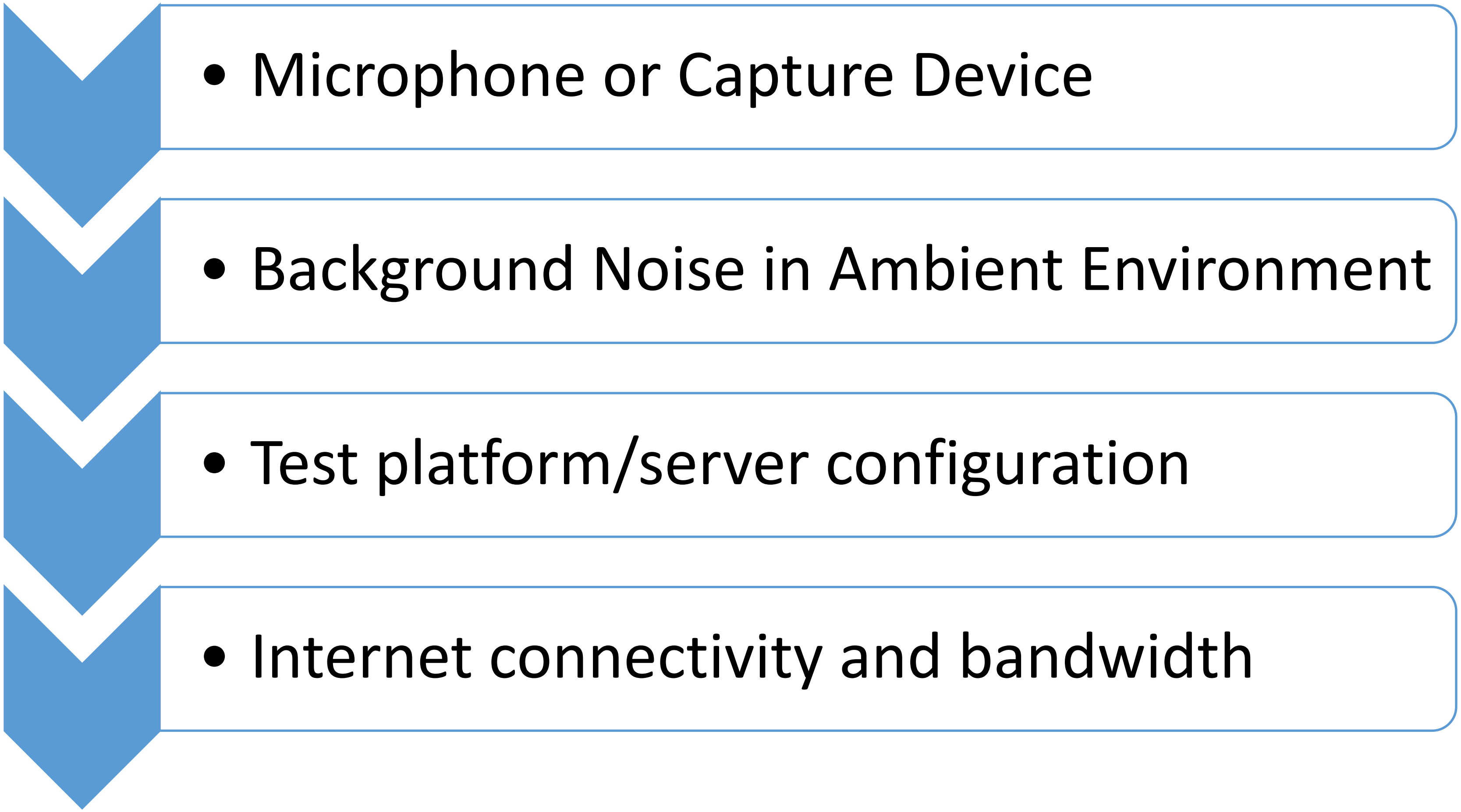
- A **rubric** is typically an evaluation tool or set of guidelines used to promote the consistent application of learning expectations, learning objectives, or learning standards in the classroom, or to measure their attainment against a consistent set of criteria.

- Elements of a scoring rubric:
  - One or more traits or dimensions that serve as the basis for judging the student response
  - Definitions and examples to clarify the meaning of each trait or dimension
  - A scale of values on which to rate each dimension
  - Standards of excellence for specified performance levels accompanied by models or examples of each level

What are the issues we encounter in non-native speech?

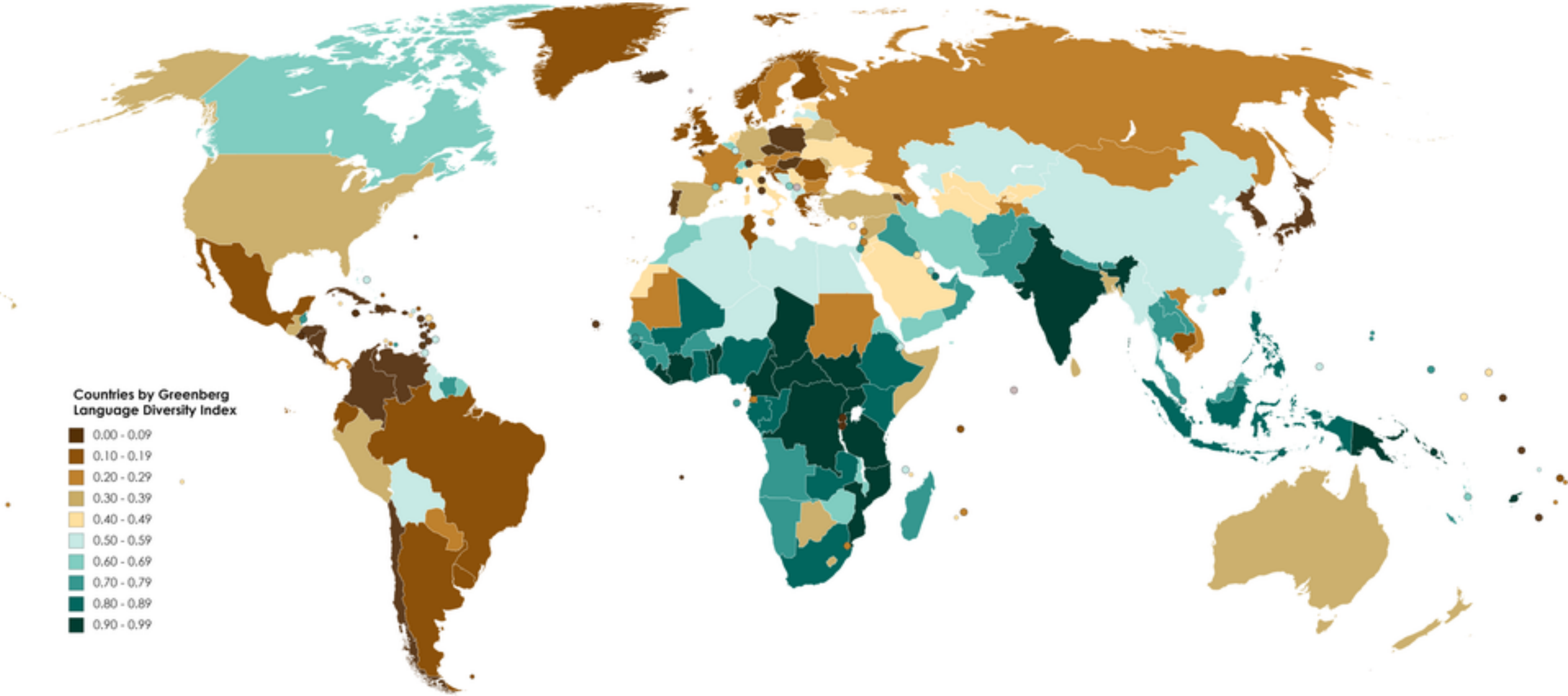
What kind of knowledge do we need to incorporate into models of non-native speech processing?

ACOUSTICS / ENVIRONMENT



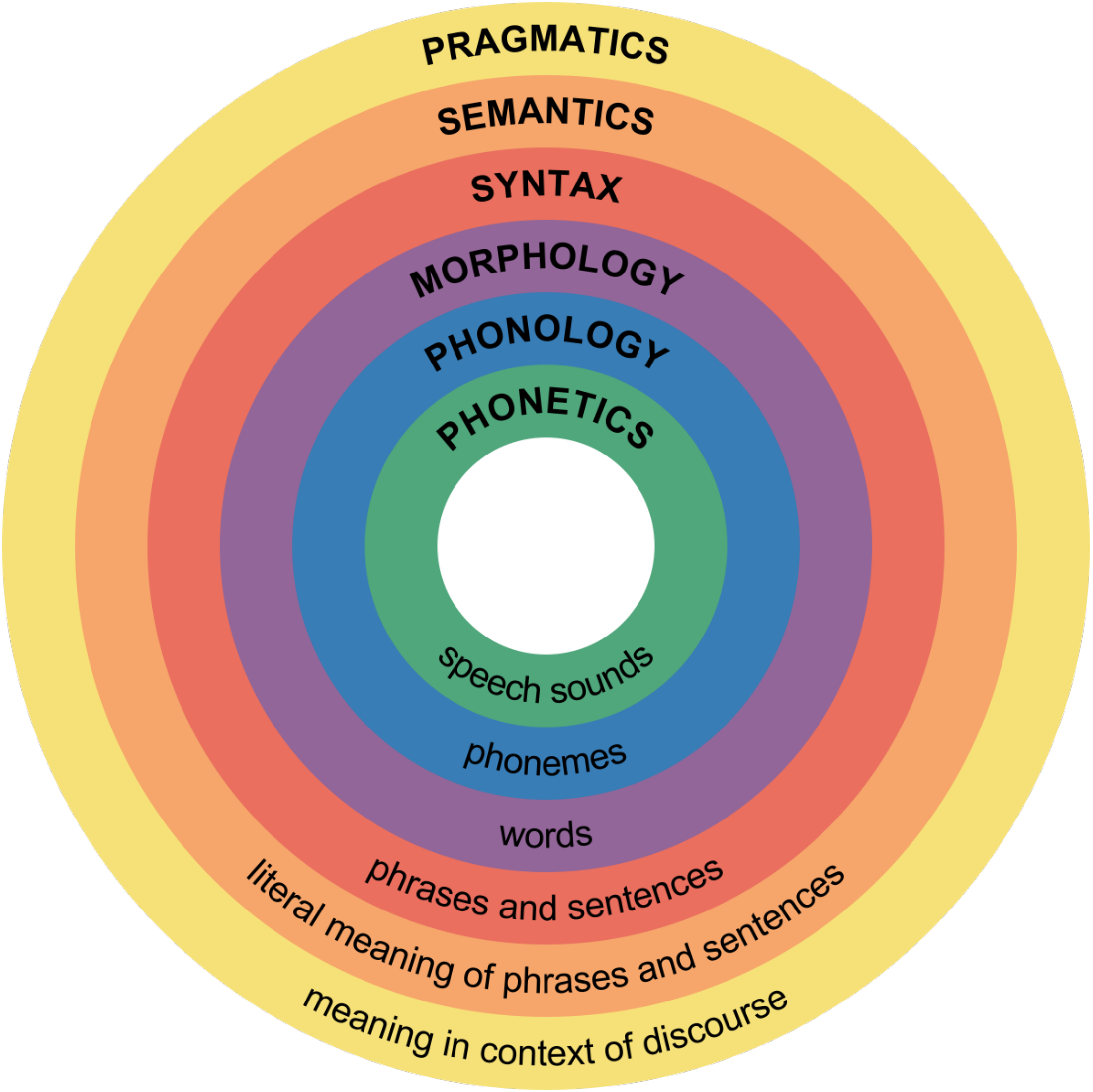
- Microphone or Capture Device
- Background Noise in Ambient Environment
- Test platform/server configuration
- Internet connectivity and bandwidth

NATIVE LANGUAGE (L1) & GEOGRAPHY



AGE & GENDER

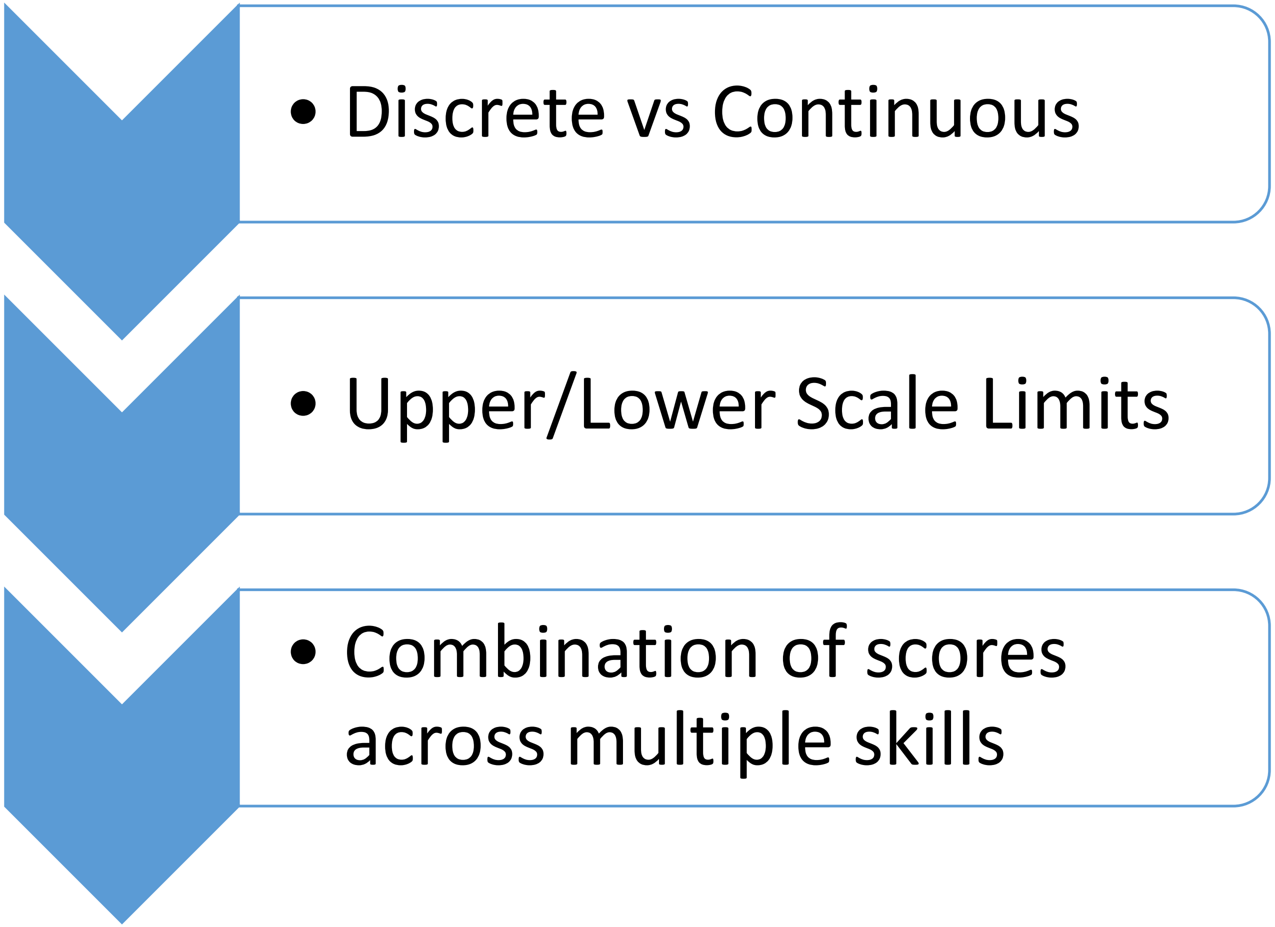
LANGUAGE & SPEECH FEATURES OF INTEREST



TASK DESIGN

- Monolog vs Dialog
- Response Duration
- Read vs Spontaneous
- Skills Measured

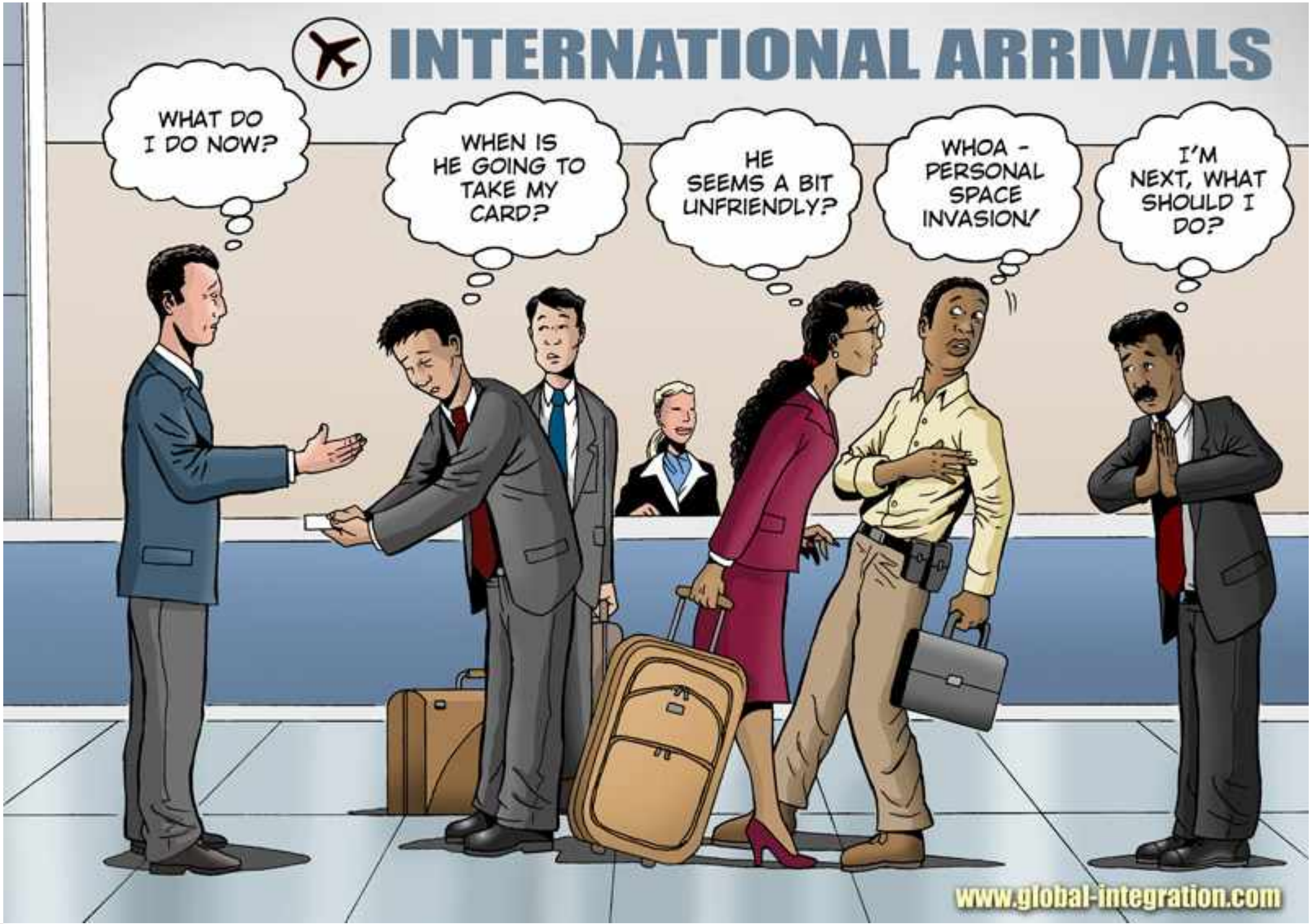
SCORE TYPE



- Discrete vs Continuous
- Upper/Lower Scale Limits
- Combination of scores across multiple skills

All these impact the type of machine learning algorithm used for automated scoring

CULTURE & PRAGMATICS


INTERNATIONAL ARRIVALS
WHAT DO I DO NOW?
WHEN IS HE GOING TO TAKE MY CARD?
HE SEEMS A BIT UNFRIENDLY?
WHOA - PERSONAL SPACE INVASION!
I'M NEXT, WHAT SHOULD I DO?
www.global-integration.com

Now that we have an understanding of the basics, let’s do a deep dive into the current state of the art in automated assessment of monolog speech…

Automatic Speech Scoring

Klaus Zechner & Keelan Evanini

10/26/20

Measuring the Power of Learning.™

AUTOMATED SCORING EXAMPLE: The lifecycle of a TOEFL iBT response

- Automated speech scoring is embedded in a much broader context beyond the NLP & Speech processing work
- Many other aspects of the lifecycle of a spoken response need to be considered
- Before the test:
  - **Test Development**: assessment specialists use Evidence Centered Design to develop construct-relevant test questions and scoring rubrics for the domain of academic English speaking proficiency
  - **Institutional Mandates**: institutions that need to evaluate students' English proficiency (e.g., universities) set requirements for test cut scores for admission based on validity studies
  - **Test Registration**: the prospective test-taker registers to take the test at a specific time and location

AUTOMATED SCORING EXAMPLE: The lifecycle of a TOEFL iBT response

- During the test:
  - **Audio capture logistics**: standardization of computer hardware/software, microphones, and testing environment to ensure high-quality audio with minimal background noise
  - **Test security**: screening mechanisms such as statistical models of responses and biometrics checks to detect fraudulent behavior (copying responses, test-taker impersonation)
- After the test:
  - **Scoring**: spoken response is sent to distributed network of human raters and automated speech scoring system; scores are combined in hybrid approach using pre-determined weights
  - **Scaling**: scores for individual test questions are aggregated for the Speaking section and scaled to a standard range (0-30) using psychometric models
  - **Score Reporting**: scores are presented in an interpretable manner on a score report and sent to test-takers and institutions

AUTOMATIC SCORING OF SPEECH

- determine speaking proficiency of a language learner
- mapping a speech recording to a numeric score
- questions:
  - how is speaking proficiency defined? how many levels are differentiated?
  - what tasks are provided to the test taker? (e.g., state an opinion, read a paragraph aloud)
  - what aspects of speaking proficiency are measured and how?
  - what is needed to train and evaluate the system?

BRIEF HISTORY

- earliest systems around 1990 (Bernstein et al., 1990): focus is on pronunciation, fluency; tasks are predictable speech (read-aloud, listen-repeat)
- early explorations into scoring of open-ended, spontaneous speech in the 2000s (Cucchiarini et al., 2002; Zechner et al., 2009)
- first operational speech scoring system for spontaneous speech: ETS's SpeechRater (2006), used for scoring TOEFL Practice Online spoken responses
- 2019: SpeechRater used for TOEFL iBT in a contributory scoring approach (combination of human and machine scores)

DIMENSIONS OF AUTOMATIC SCORING

Dimension	Easier to score	Harder to score
Speech predictability	Highly predictable (e.g., read aloud)	Highly unpredictable (e.g., spontaneous speech)
Response based on stimulus/source	Tightly connected to stimulus materials (e.g., prompts with lectures, images, and readings)	Not connected to particular stimulus materials (e.g., prompts ask for personal knowledge, experience, or opinions)
Task interactivity	Monologic	Dialogic

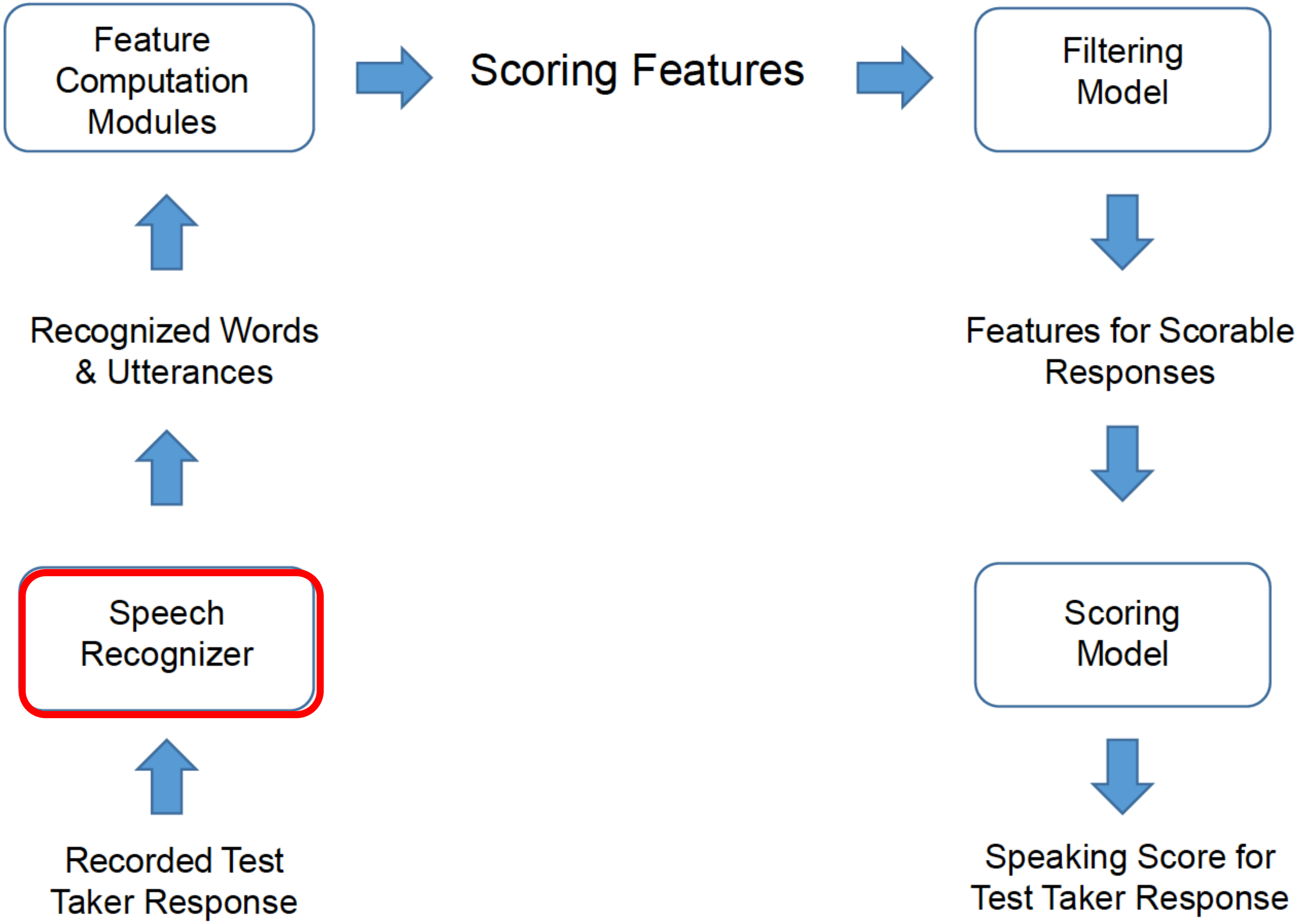
DIMENSIONS OF AUTOMATIC SCORING

Dimension	Easier to score	Harder to score
Test taker native languages	Only one or very few native languages spoken	Many different native languages spoken
Test taker age	Adults	Children
Test taker proficiency levels	Mix of proficiencies	Fairly similar proficiency

DIMENSIONS OF AUTOMATIC SCORING

Dimension	Easier to score	Harder to score
Speech capture	Professional recording equipment such as close-talk noise-cancelling microphone	Consumer devices such as mobile device with built-in microphone
Overall recording environment	Low noise (e.g., quiet room, soundproof recording booth)	High and variable noise (e.g., café, street, classroom)
Voice contamination during recording	Only single speaker recorded	Potentially multiple speakers around or in background

SPEECHRATER SYSTEM ARCHITECTURE



Recorded Test Taker Response

Speech Recognizer

Recognized Words & Utterances

Feature Computation Modules

Scoring Features

Filtering Model

Features for Scorable Responses

Scoring Model

Speaking Score for Test Taker Response

CHALLENGES FOR ASR OF SPONTANEOUS NON-NATIVE SPEECH

- Acoustic Modelling
  - high variability of phones’ acoustics across many different first languages
  - variation also due to range of speaking proficiency
  - Mispronunciations

- Language Modelling
  - lexical and grammatical errors
  - more and less regular filled pauses, other disfluencies

AUTOMATIC SPEECH RECOGNITION COMPONENT IN SPEECHRATER

- Kaldi, nnet2
- AM: using i-vectors
- LM: trigram model
- training corpus: ~800 hours of spontaneous non-native speech
- WER ~20%
- agreement between human transcribers: 15% - 20% WER

SPEAKING PROFICIENCY & TASKS

- most important question when designing an assessment: what do we want to measure?
- speaking proficiency – in TOEFL iBT: communicative competence
- domain: academic, social/campus
- speaking tasks: need to elicit speech from test takers that allow to evaluate the aspects of speaking proficiency that are considered to be relevant
- informed by theories of language learning and second language acquisition
- TOEFL iBT: independent and integrated tasks

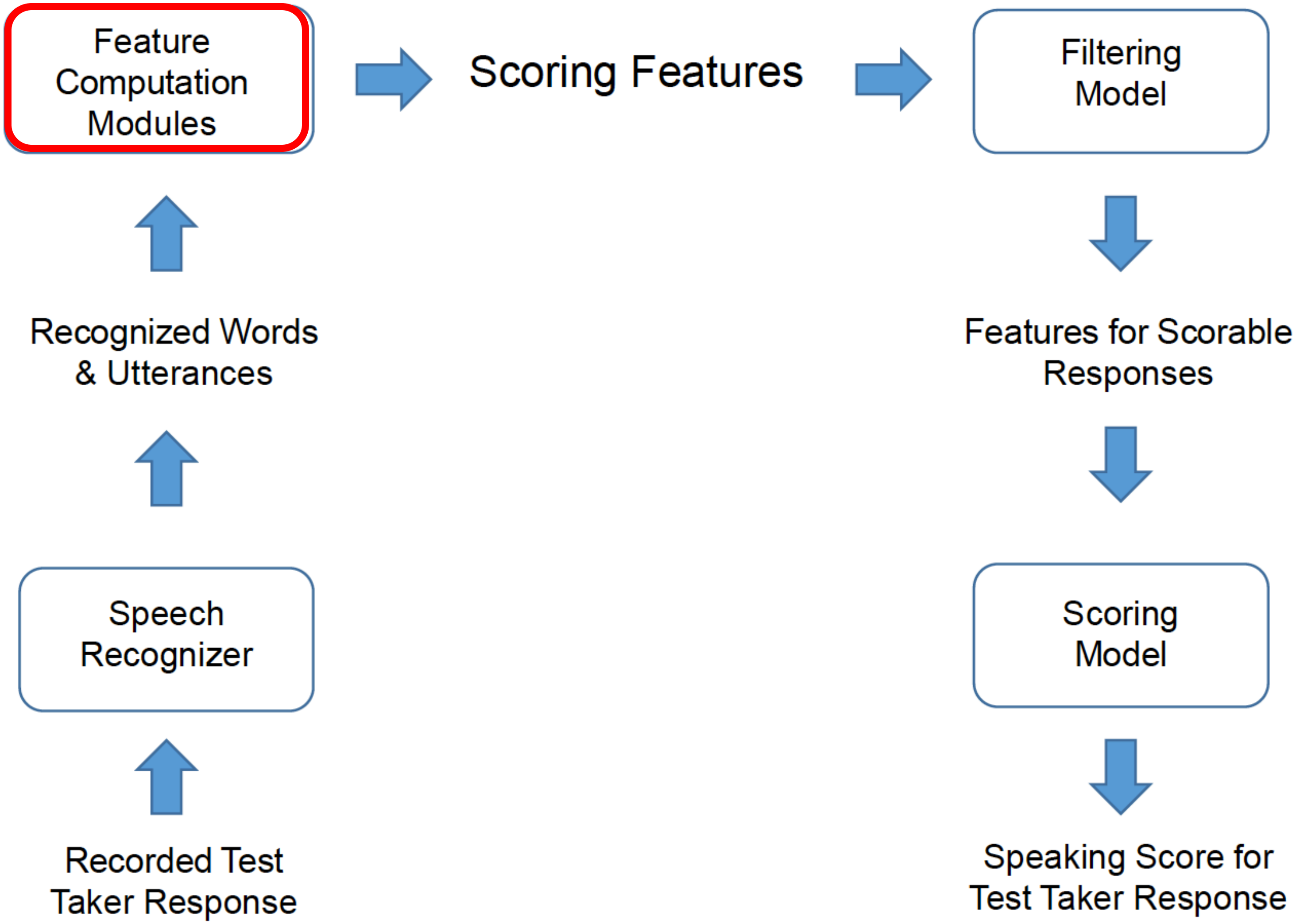
SPEAKING RUBRICS

- detailed descriptions of important aspects of speaking proficiency for each score level

- used by human raters to assign holistic scores to spoken responses

- TOEFL iBT: 4 score levels, 3 dimensions: delivery (fluency, pronunciation), language use (vocabulary, grammar), topic development (content, discourse)

SPEECHRATER SYSTEM ARCHITECTURE



Feature Computation Modules

Scoring Features

Filtering Model

Recognized Words & Utterances

Features for Scorable Responses

Speech Recognizer

Scoring Model

Recorded Test Taker Response

Speaking Score for Test Taker Response

FEATURES MEASURING SPEAKING PROFICIENCY

- using information from speech processing components (e.g., F0 or power obtained via Praat), word, timing and confidence information from ASR, phone and syllable boundaries from forced alignment
- some features require complex processing using NLP methods, e.g., predicting stressed syllables, syntactic parsing etc.
- around 80 features computed in SpeechRater for TOEFL iBT

DELIVERY:
FLUENCY & PRONUNCIATION

- Fluency: speaking rate, distribution of pauses, disfluencies
  - example: speaking_rate = #words/response_length
  - example: pause_variability = stddev(pause_lengths)
- Pronunciation: segmental (individual phones), suprasegmental (stress, pitch contours, rhythm etc.)
  - example: pronunciation_accuracy = average AM score across all phones (where AM is based on native speech, used in forced alignment)
  - example: stress_frequency = percentage of syllables bearing stress

LANGUAGE USE: VOCABULARY & GRAMMAR

- Vocabulary: diversity, range (e.g., use of frequent vs. infrequent words)
  - example: vocabulary_sophistication = average frequency of words in response based on a word list

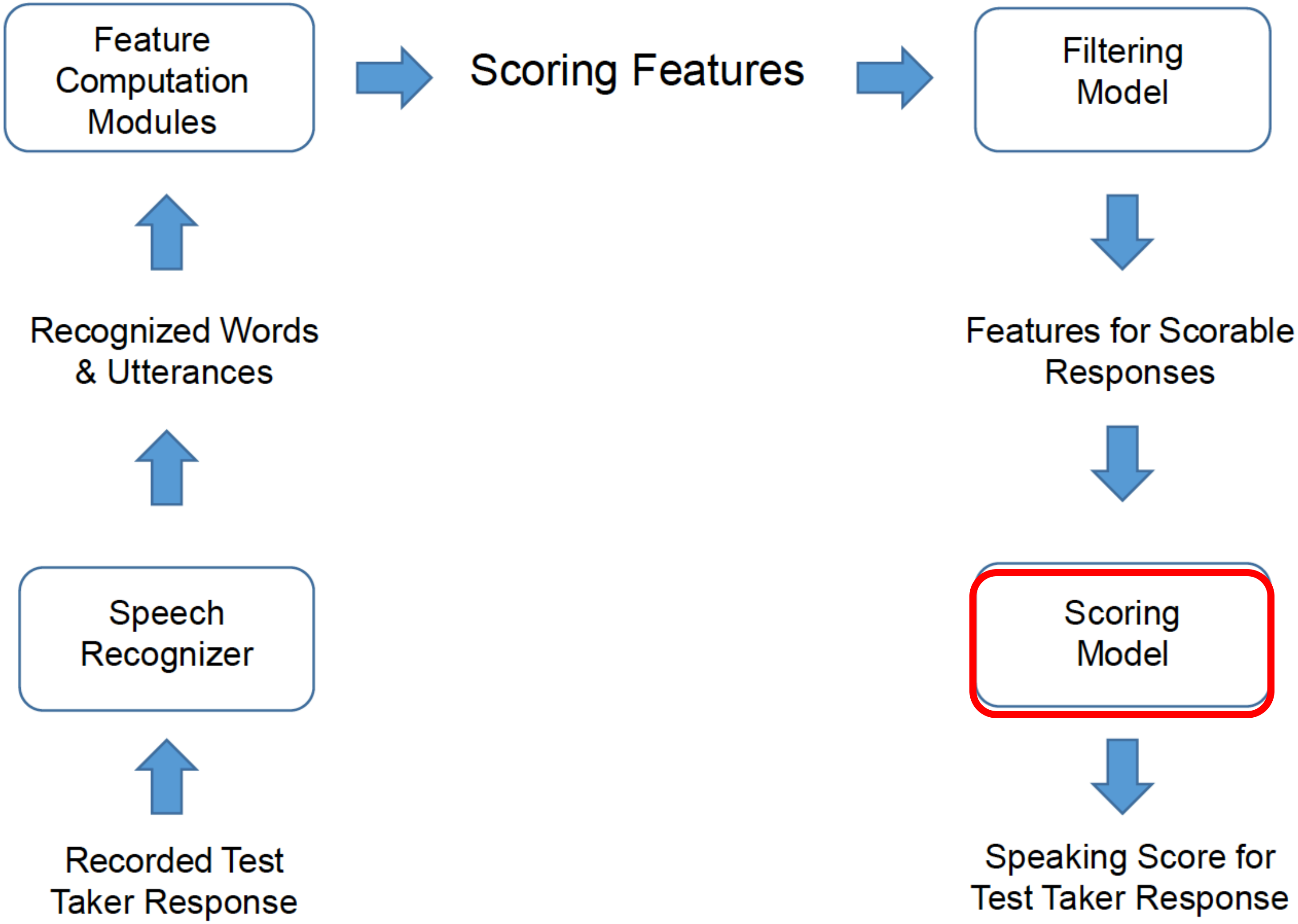
- Grammar: accuracy (measured via POS n-grams), diversity (occurrence of syntactic constructions)
  - example: complexity of phrases = number of noun phrases with embeddings

TOPIC DEVELOPMENT: CONTENT & DISCOURSE

- Content: bag-of-words features (general topic), keypoint prediction (specific content in a response)
  - example: topicality = cosine_similarity (words in response, words in high proficient responses)

- Discourse: coherence/cohesion, structure (using RST parsing)
  - example: coherence = # connection words (e.g., “and”, “then”, ...)

SPEECHRATER SYSTEM ARCHITECTURE



Feature Computation Modules

Scoring Features

Filtering Model

Recognized Words & Utterances

Features for Scorable Responses

Speech Recognizer

Scoring Model

Recorded Test Taker Response

Speaking Score for Test Taker Response

CONDUCT PILOT TEST & OBTAIN HUMAN RATINGS

- recommended: 1000+ responses for each task, double human scored

- used operational data: 200k test takers, 50% training, 50% evaluation

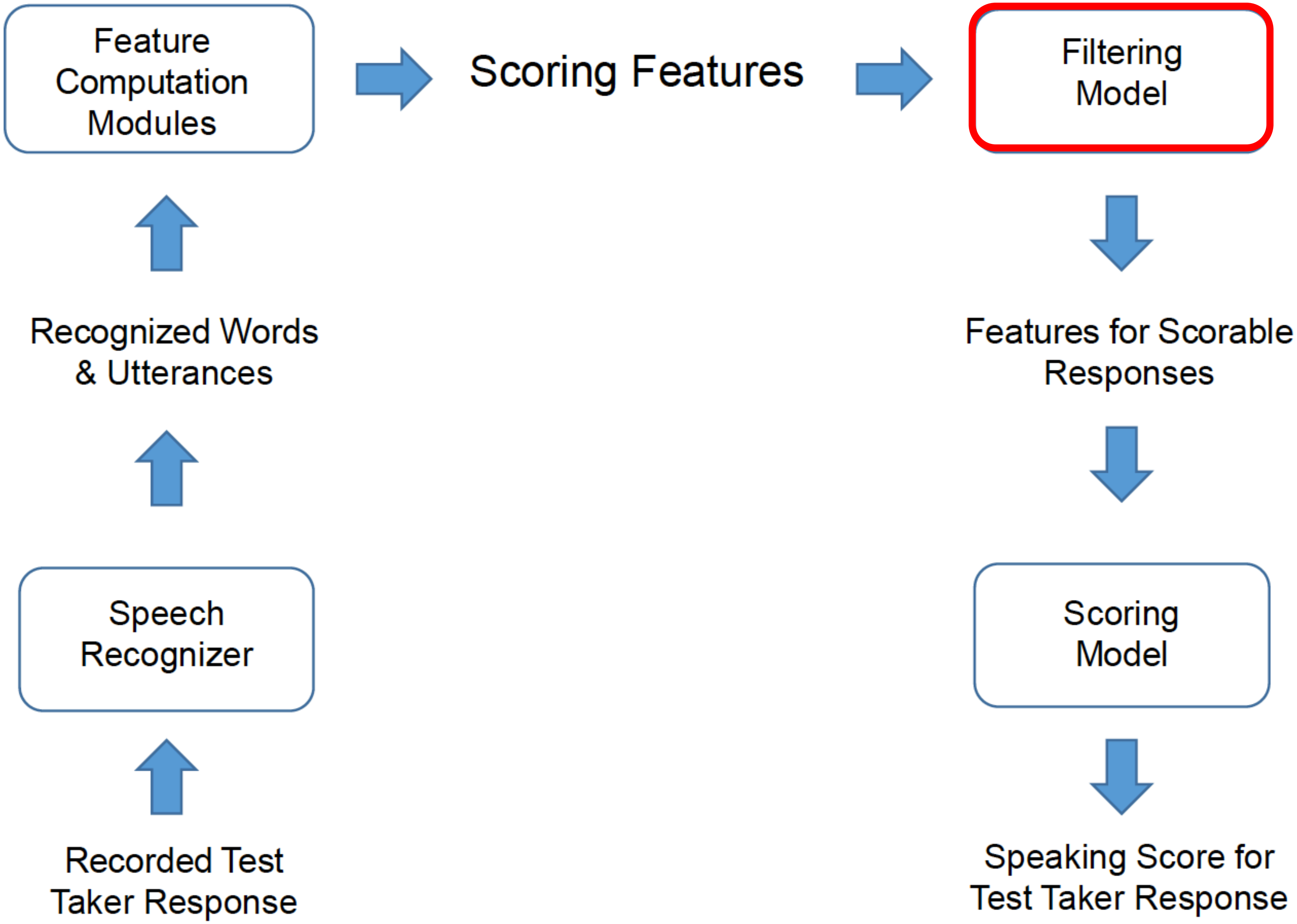
- 4 responses per test taker

- human ratings: 1-4, integer

SCORING MODEL

- reduce feature set via non-negative least-square regression (avoid high inter-correlations between features)
- 28 features, from all areas of the speaking rubric (except for “content’)
- correlation between machine and human scores: 0.64 (*inter-human correlation: 0.58*)
- using other traditional machine learning algorithms did not result in any marked performance increase
- recently explored transformer-based scoring, more promising (but disadvantage of less interpretable scores)

SPEECHRATER SYSTEM ARCHITECTURE



Recorded Test Taker Response

Speech Recognizer

Recognized Words & Utterances

Feature Computation Modules

Scoring Features

Filtering Model

Features for Scorable Responses

Scoring Model

Speaking Score for Test Taker Response

FILTERING MODELS

- some spoken responses can have non-ideal properties: blank, high-noise, off-topic...
- need to identify such non-scorable responses: filtering models
- TOEFL iBT: filtering model based on acoustic information (e.g., noisy or empty responses)
- flagged responses are not scored by the machine but typically routed to human raters

SYSTEM EVALUATION

- major evaluation criteria, informed by theories of educational measurement and psychometrics, are: reliability, validity, and fairness
- reliability: machine score should correspond to the “true score” (idealized score e.g. average of many independent human raters’ scores)
- validity: are features measuring what they are supposed to measure? (in terms of the relevant aspects of speaking proficiency)
- fairness: do automated scores show any bias towards sub-groups, e.g., gender, native language etc.
- TOEFL iBT: hybrid scoring system: for each item, a machine score and a human rater score are combined
- combines advantages of human raters (evaluation of content aspects) and machine (evaluation of more detailed aspects of fluency, pronunciation etc.)

DEEP LEARNING FOR AUTOMATED SCORING

- using BLSTMs for scoring of spontaneous speech, using both time-aggregated and response-level linguistic features (Yu et al., 2015)

- using BLSTMs for scoring of dialogs and monologs, predicting both holistic and analytic scores (Qian et al., 2019)

- using transformers to score specific content in spontaneous speech (Wang et al., 2020)

LARGE SCALE DEPLOYMENT: ISSUES AND CONSIDERATIONS

- test administration: test centers vs. at-home
- bandwidth for A/V data (video e.g. used for test security)
- test integrity/security issues
- data flow and storage
- integration of human raters and machine scores
- score turn-around time to test takers
- monitoring of scores (human and machine)
- continuous supply of new items

MAJOR DIRECTIONS MOVING FORWARD

- expanding feature set (e.g., discourse structure, specific content)

- exploring DL for scoring

- expanding set of filtering models

- improving ASR component (e.g., using larger data set for training)

Automatic Feedback

Keelan Evanini

10/26/20

Measuring the Power of Learning.™

Automated Speech Feedback

- Goal: provide language learners with actionable information about how they can improve their speaking ability
- Focus is on spontaneous speech in an academic context
  - Prior research on speech feedback has primarily focused on restricted speech, e.g., pronunciation error detection
- Considerations
  - Which aspects of speaking proficiency to provide feedback on?
  - How should feedback be presented?

Feedback Feature Selection Criteria

Usability	Understandable to test takers
Usefulness	Provide actionable information
Reliability	Be consistent
Relevance	Correlate with human scores
Coverage	Address different aspects of the construct

Selected SpeechRater Features

Feature name	Construct area	Description
Speaking Rate	Delivery-Fluency	Words per second
Sustained Speech	Delivery-Fluency	Number of words without disfluencies
Pause Frequency	Delivery-Fluency	Pauses per word
Repetitions	Delivery-Fluency	Number of repetitions
Vowels	Delivery-Pronunciation	Vowel sounds compared to a native speaker model
Rhythm	Delivery-Pronunciation	Stressed syllables
Vocabulary depth	Language Use-Vocabulary	Use of infrequent words

Consistency across Tasks

TPO 2012 (N = 776)	Avg. Alpha	Min. Alpha	Alpha Range
Speaking Rate	0.94	0.91	0.04
Sustained Speech	0.91	0.88	0.05
Pause Frequency	0.94	0.92	0.02
Repetitions	0.75	0.67	0.15
Rhythm	0.83	0.80	0.07
Vowels	0.88	0.78	0.15
Vocabulary Depth	0.79	0.74	0.09
TOEFL iBT (N = 10,469)	**Avg. Alpha**	**Min. Alpha**	**Alpha Range**
Speaking Rate	0.95	0.94	0.02
Sustained Speech	0.92	0.86	0.07
Pause Frequency	0.95	0.94	0.02
Repetitions	0.80	0.79	0.04
Rhythm	0.85	0.83	0.05
Vowels	0.81	0.66	0.23
Vocabulary Depth	0.84	0.81	0.07

Consistency across Tasks

TPO 2012 (N = 776)	Avg. Alpha	Min. Alpha	Alpha Range
Speaking Rate	0.94	0.91	0.04
Sustained Speech	0.91	0.88	0.05
Pause Frequency	0.94	0.92	0.02
Repetiti			0.15
Rhythm			0.07
Vowels	0.88	0.78	0.15
Vocabulary Depth	0.79	0.74	0.09

TPO Average Cronbach's alpha: 0.75 – 0.94

TOEFL iBT (N = 10,469)	Avg. Alpha	Min. Alpha	Alpha Range
Speaking Rate	0.95	0.94	0.02
Sustained Speech	0.92	0.86	0.07
Pause Frequency	0.95	0.94	0.02
Repetiti			
Rhythm			
Vowels	0.81	0.66	0.23
Vocabulary Depth	0.84	0.81	0.07

TOEFL Average Cronbach's alpha: 0.80 – 0.95

Correlation with Human Score

Name for reporting	Mean Corr. (TPO N = 776)	Mean Corr. (TOEFL iBT N = 10,469)
Speaking Rate	0.51	0.69
Sustained Speech	0.50	0.62
Pause Frequency	0.50	0.66
Repetitions	0.32	0.46
Rhythm	0.45	0.61
Vowels	0.48	0.57
Vocabulary Depth	0.48	0.67

Correlation with Human Score

Name for reporting	Mean Corr. (TPO N = 776)	Mean Corr. (TOEFL iBT N = 10,469)
Speaking Rate	0.51	0.69
Sustained S		
Pause Frequency	0.50	0.66
Repetitions	0.32	0.46
Rhythm		
Vowels	0.48	0.57
Vocabulary Depth	0.48	0.67

TPO Pearson correlations: 0.32 – 0.51

TOEFL Pearson correlations: 0.46 – 0.69

Added Value

- Do machine-generated linguistic measures have value in addition to human and machine holistic scores?
- Approach: Proportional Reduction in Mean Square Error (PRMSE) (Haberman, 2008; Sinharay, Puhan, & Haberman, 2011)

PRMSE = Proportion of the variance explained by the observed scores to the total variance of true scores.

- The larger the PRMSE, the better the model prediction.

➢ PRMSE addresses relationship between observed and true scores, as well as reliability of observed measures.

In our case, feature scores have added value when:

$$\mathrm{PRMSE}_{\mathrm{feature\_score}} > \mathrm{PRMSE}_{\mathrm{holistic\_score}}$$

Added Value - Model Specification

Best linear prediction approach (BLP) to predict true scores from observed scores.

- Model 1: Observed **feature score** used to predict the corresponding feature true score.
- Model 2: Observed **machine holistic score** to predict the feature true score.
- Model 3: Observed **human holistic score** to predict the feature true score.

Added Value - Modeling Results

(*N*=244)	PRMSE		
Dependent	Model 1 AS-Feature	Model 2 AS-Holistic	Model 3 Human Holistic
Sustained speech	.92	.65	.38
Pause frequency	.94	.60	.33
Speaking rate	.94	.70	.36
Repetitions	.78	.21	.07
Vowels	.92	.58	.33
Rhythm	.83	.55	.30
Vocabulary depth	.79	.68	.44

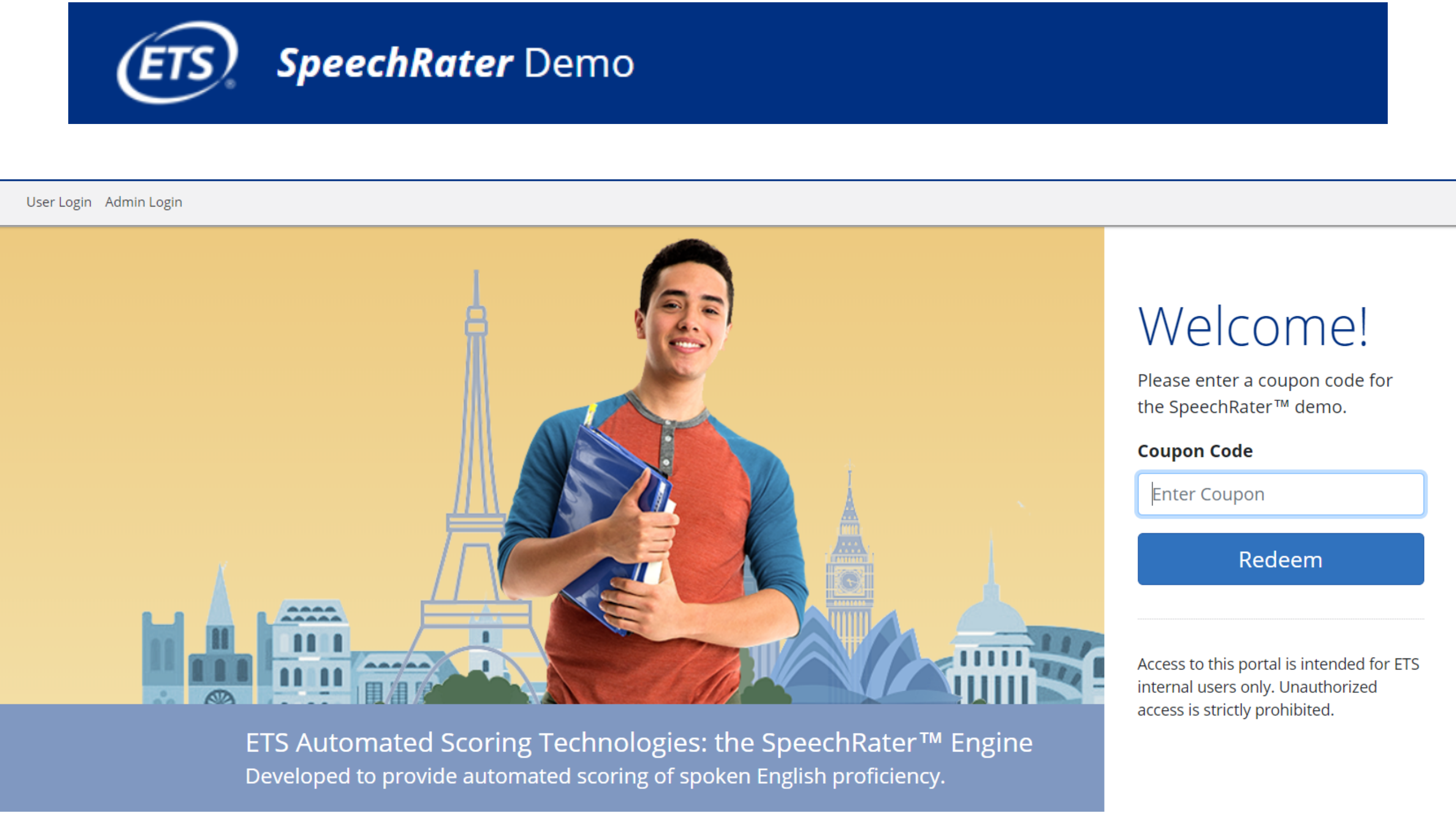
Substantial gain in PRMSE between Model 1 and 2, and between Model 1 and 3

Added Value - Modeling Results

(*N*=244)	PRMSE		
	Model 1	Model 2	Model 3
Dep			ic
Su			
Pa			
Sp			
Re			
Vowels	.92	.58	.33
Rhythm	.83	.55	.30
Vocabulary depth	.79	.68	.44

Feature scores have added value over holistic scores because feature true scores were better predicted by the corresponding observed feature scores than by machine or human holistic score.

Substantial gain in PRMSE between Model 1 and 2, and between Model 1 and 3



SpeechRater Demo

Welcome!

Please enter a coupon code for the SpeechRater™ demo.

Coupon Code

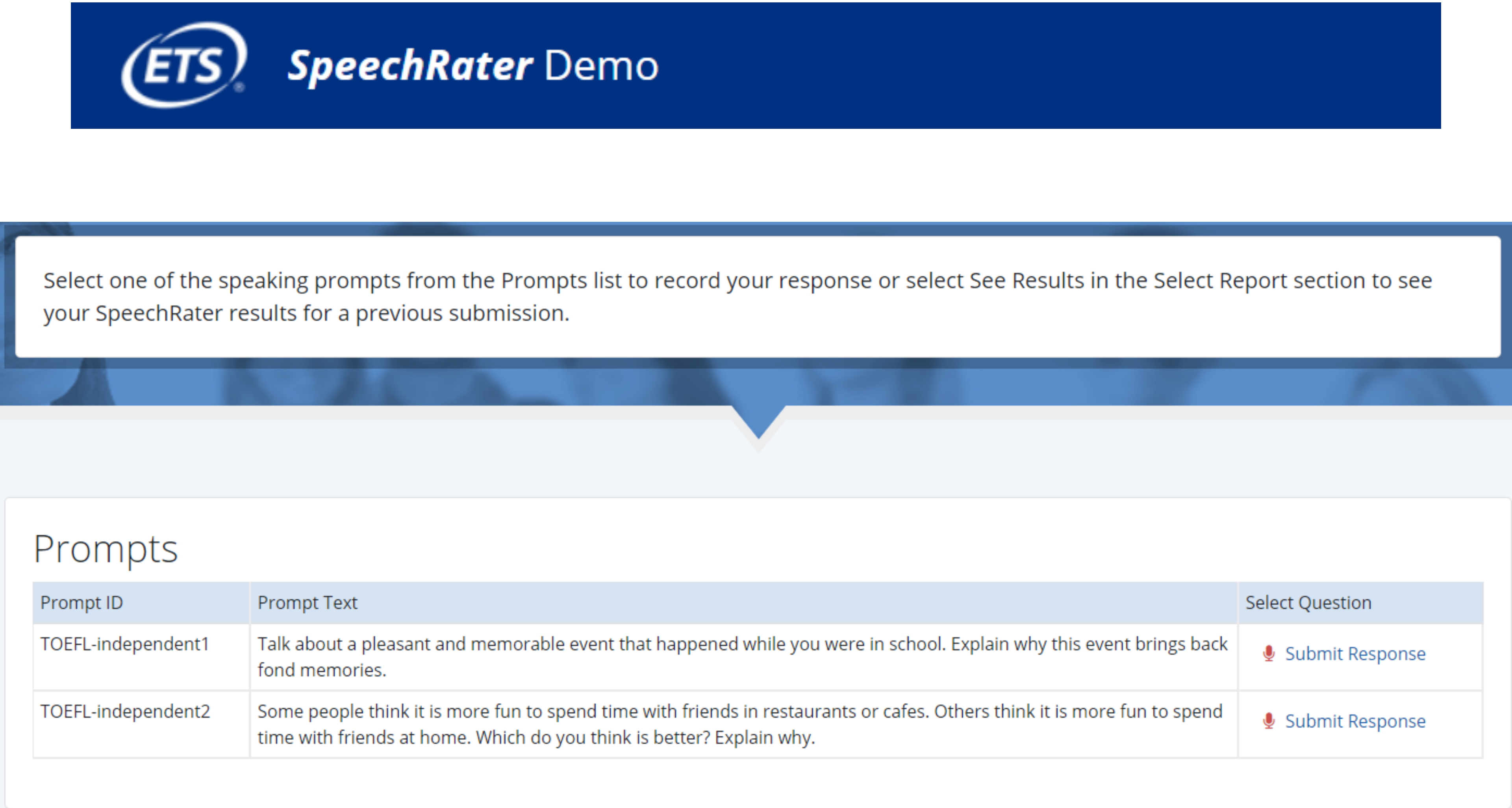
Enter Coupon

Access to this portal is intended for ETS internal users only. Unauthorized access is strictly prohibited.

ETS Automated Scoring Technologies: the SpeechRater™ Engine

Developed to provide automated scoring of spoken English proficiency.

https://nlp-pilot.ets.org/srater

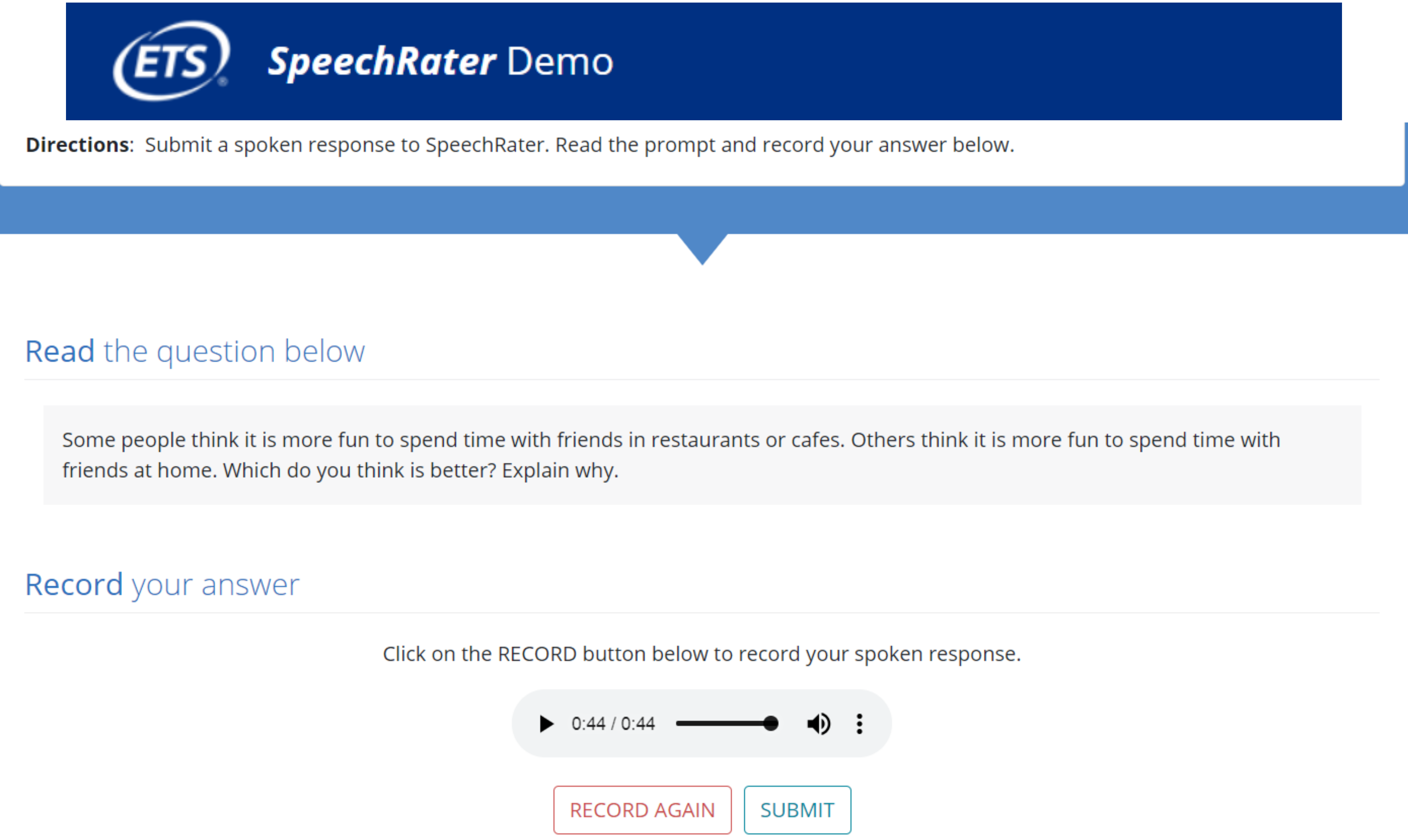


SpeechRater Demo

Select one of the speaking prompts from the Prompts list to record your response or select See Results in the Select Report section to see your SpeechRater results for a previous submission.

Prompts

Prompt ID	Prompt Text	Select Question
TOEFL-independent1	Talk about a pleasant and memorable event that happened while you were in school. Explain why this event brings back fond memories.	Submit Response
TOEFL-independent2	Some people think it is more fun to spend time with friends in restaurants or cafes. Others think it is more fun to spend time with friends at home. Which do you think is better? Explain why.	Submit Response



SpeechRater Demo

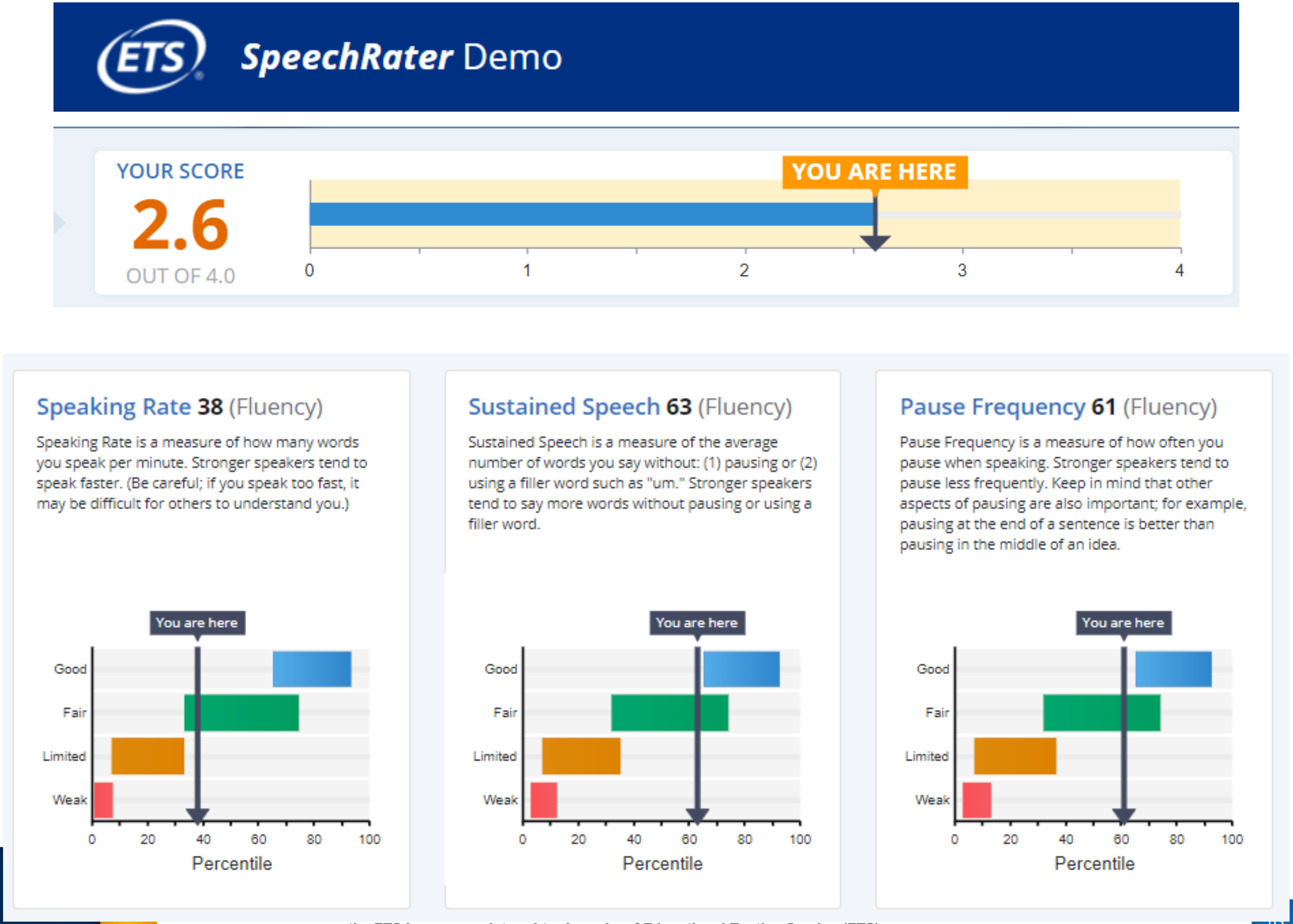
Directions: Submit a spoken response to SpeechRater. Read the prompt and record your answer below.

Read the question below

Some people think it is more fun to spend time with friends in restaurants or cafes. Others think it is more fun to spend time with friends at home. Which do you think is better? Explain why.

Record your answer

Click on the RECORD button below to record your spoken response.



ETS SpeechRater Demo

YOUR SCORE

2.6

OUT OF 4.0

Speaking Rate **38** (Fluency)

Speaking Rate is a measure of how many words you speak per minute. Stronger speakers tend to speak faster. (Be careful; if you speak too fast, it may be difficult for others to understand you.)

Sustained Speech **63** (Fluency)

Sustained Speech is a measure of the average number of words you say without: (1) pausing or (2) using a filler word such as "um." Stronger speakers tend to say more words without pausing or using a filler word.

Pause Frequency **61** (Fluency)

Pause Frequency is a measure of how often you pause when speaking. Stronger speakers tend to pause less frequently. Keep in mind that other aspects of pausing are also important; for example, pausing at the end of a sentence is better than pausing in the middle of an idea.

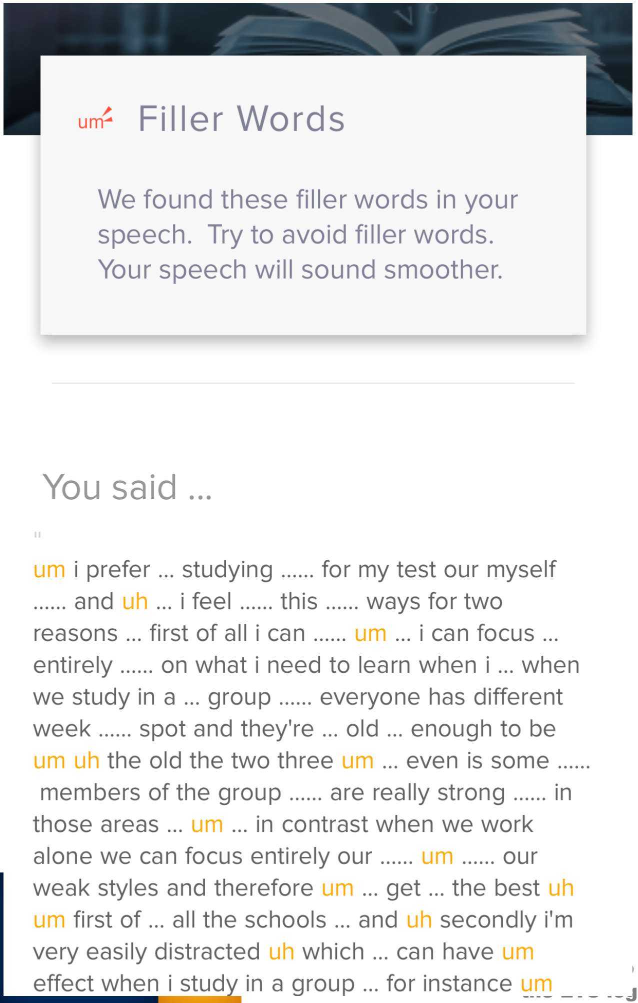
User Perception Study (Gu et al., 2020)

- Participants: 123 TPO test takers, 31 EFL teachers
- Review a mockup report for an imaginary test taker
- Survey: Rate the usefulness of the feedback for
  - Preparing for TOEFL iBT
  - Improving English speaking proficiency in general
- Four-point Likert-scale: “definitely yes”, “maybe yes”, “maybe no”, and “definitely no”
  - Table reports the percentage of the respondents who said “definitely yes” or “maybe yes”

	For Test Prep	For Improvement
Test Takers	92%	86%
Teachers	87%	80%

Targeted Feedback

- In addition to global feedback about an entire response, *targeted feedback* can also be helpful to learners



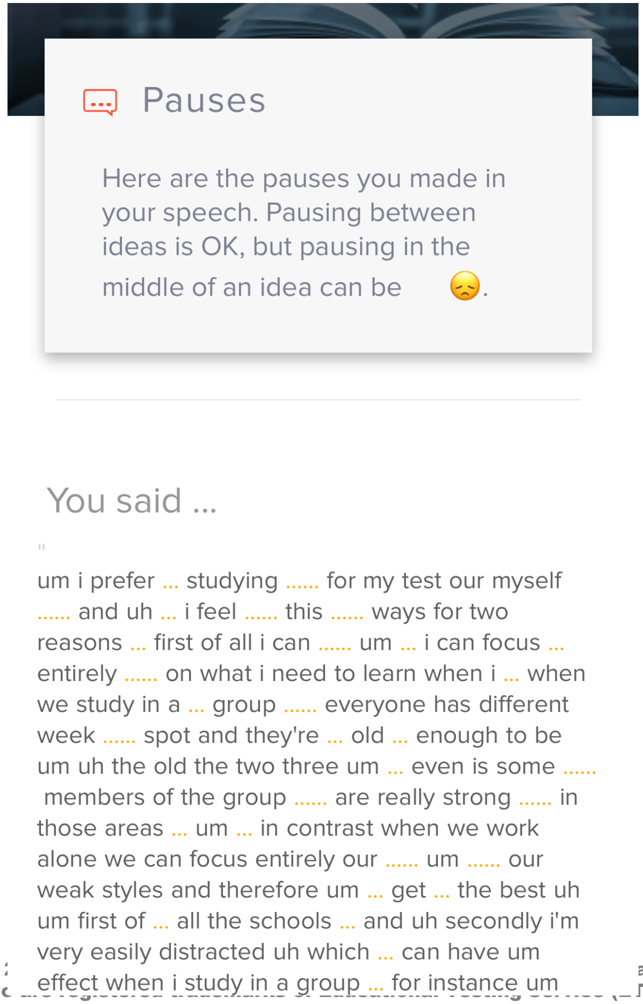
Filler Words

We found these filler words in your speech. Try to avoid filler words. Your speech will sound smoother.

You said ...

"

um i prefer ... studying ...... for my test our myself ...... and uh ... i feel ...... this ...... ways for two reasons ... first of all i can ...... um ... i can focus ... entirely ...... on what i need to learn when i ... when we study in a ... group ...... everyone has different week ...... spot and they're ... old ... enough to be um uh the old the two three um ... even is some ...... members of the group ...... are really strong ...... in those areas ... um ... in contrast when we work alone we can focus entirely our ...... um ...... our weak styles and therefore um ... get ... the best uh um first of ... all the schools ... and uh secondly i'm very easily distracted uh which ... can have um effect when i study in a group ... for instance um



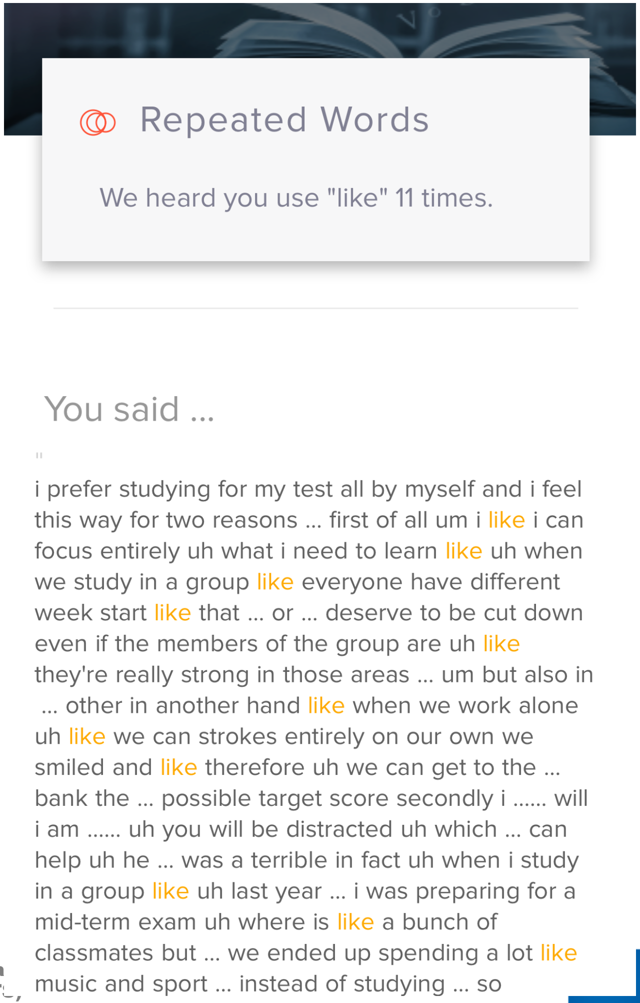
Pauses

Here are the pauses you made in your speech. Pausing between ideas is OK, but pausing in the middle of an idea can be 😞.

You said ...

"

um i prefer ... studying ...... for my test our myself ...... and uh ... i feel ...... this ...... ways for two reasons ... first of all i can ...... um ... i can focus ... entirely ...... on what i need to learn when i ... when we study in a ... group ...... everyone has different week ...... spot and they're ... old ... enough to be um uh the old the two three um ... even is some ...... members of the group ...... are really strong ...... in those areas ... um ... in contrast when we work alone we can focus entirely our ...... um ...... our weak styles and therefore um ... get ... the best uh um first of ... all the schools ... and uh secondly i'm very easily distracted uh which ... can have um effect when i study in a group ... for instance um



Repeated Words

We heard you use "like" 11 times.

You said ...

"

i prefer studying for my test all by myself and i feel this way for two reasons ... first of all um i like i can focus entirely uh what i need to learn like uh when we study in a group like everyone have different week start like that ... or ... deserve to be cut down even if the members of the group are uh like they're really strong in those areas ... um but also in ... other in another hand like when we work alone uh like we can strokes entirely on our own we smiled and like therefore uh we can get to the ... bank the ... possible target score secondly i ...... will i am ...... uh you will be distracted uh which ... can help uh he ... was a terrible in fact uh when i study in a group like uh last year ... i was preparing for a mid-term exam uh where is like a bunch of classmates but ... we ended up spending a lot like music and sport ... instead of studying ... so

PART II: TRENDING TOPICS

DIALOG & MULTIMODAL SCORING AND FEEDBACK

Spoken Dialog Systems in the Educational Domain

Vikram Ramanarayanan

10/26/20

Measuring the Power of Learning.™

WHY DIALOG?

- Current instruments for spoken formative English language assessment are **not** naturalistic and conversational, for the most part.

- Certain exceptions (or getting there at least):

supiki

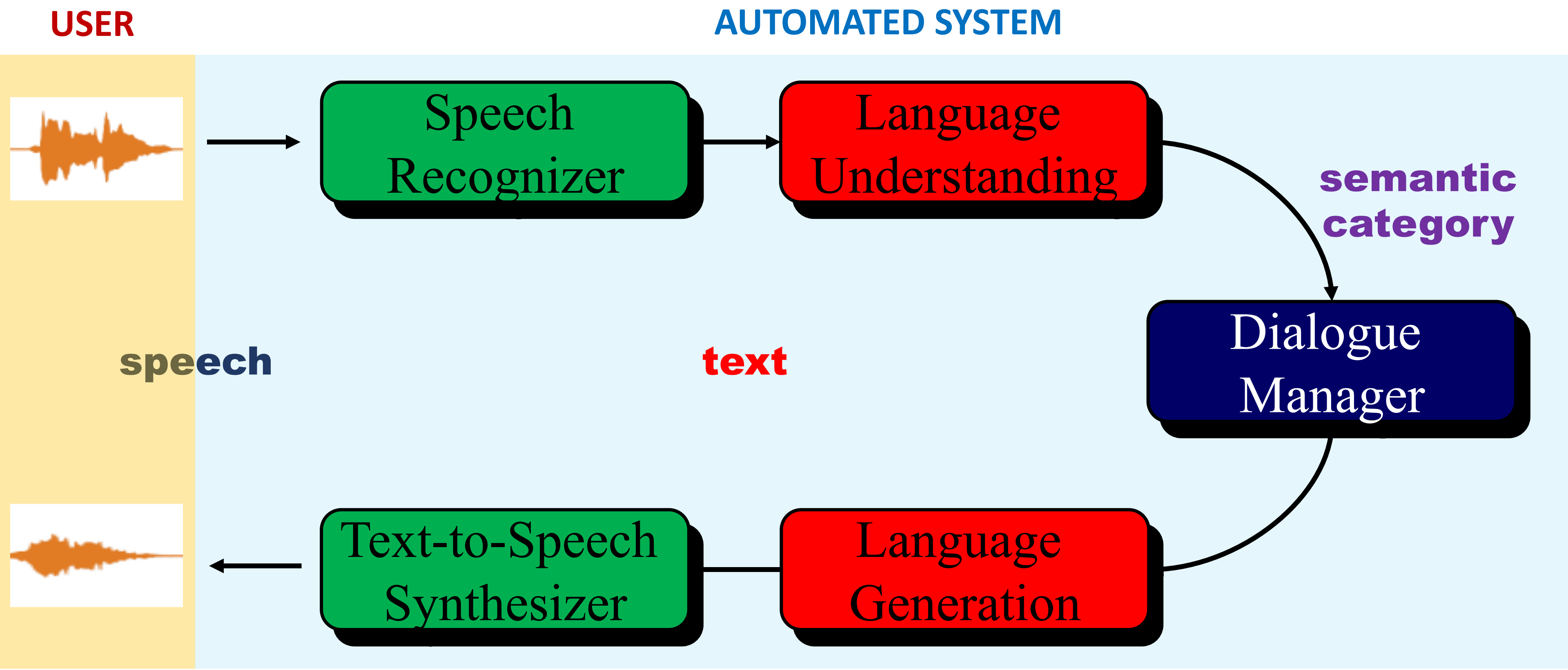
duolingo

alelo®

- However, it is important to engage people in conversations in naturalistic settings in order to be more effective and measure/train relevant skills

- Reaching people with disabilities, e.g., visually-impaired individuals.

SPOKEN DIALOG SYSTEMS



A dialog system iteratively goes through this loop until termination.

Adapted from: Kamboj, SIG-AI Fall 2003 presentation

SEVERAL CHALLENGES IN DESIGNING DIALOG SYSTEMS

Voice activity detection:

- Sensitivity
- Time-out
- Barge-in

ASR:

- Language is strongly context-dependent
- Adverse acoustic conditions (speaker phone, background noise)

SLU:

- Out-of-vocabulary/no-match

DM:

- Behave naturally (e.g. back-channel),
- robust (cover all situations, do not loop indefinitely, …),
- effective (accomplish target task with reasonable effort)

COMPLEMENTARY TUTORIALS

- Our 2018 Interspeech Tutorial on Spoken Dialog Technology for Educational Domain Applications
  - http://www.vikramr.com/tutorials.html

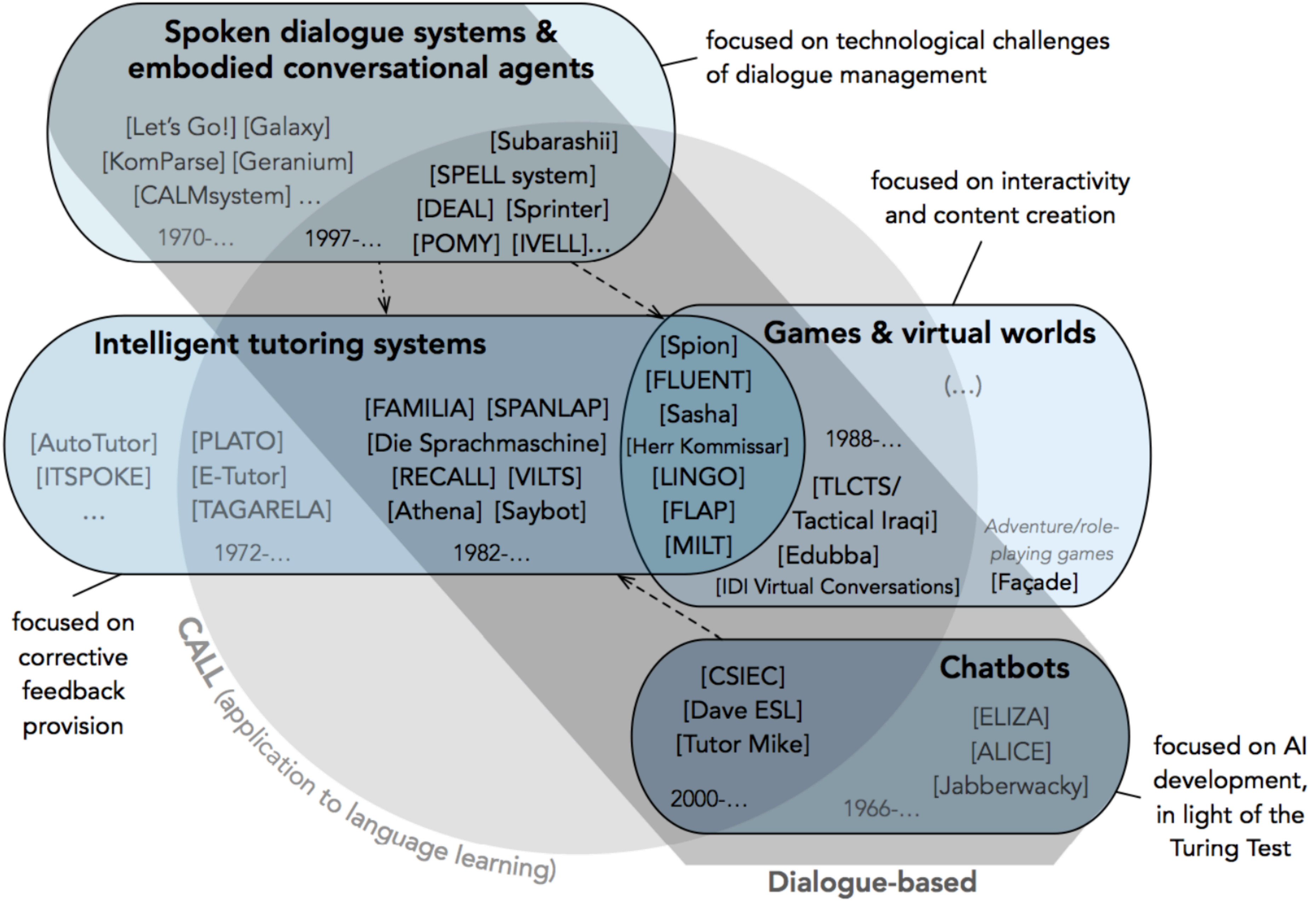
- Deep Learning for Dialog Systems by Vivian Chen, Asli Celikyilmaz and Dilek Hakkani-Tur
  - https://www.csie.ntu.edu.tw/~yvchen/doc/DeepDialogue_Tutorial.pdf

- Deep Learning for Conversational AI by Pei-Hao Su, Nikola Mrksic, Iñigo Casanueva and Ivan Vulic
  - https://www.poly-ai.com/docs/naacl18.pdf

- Tutorial material from Steve Young (Cambridge) on Statistical Spoken Dialogue
  - http://mi.eng.cam.ac.uk/~sjy/tutorial.html

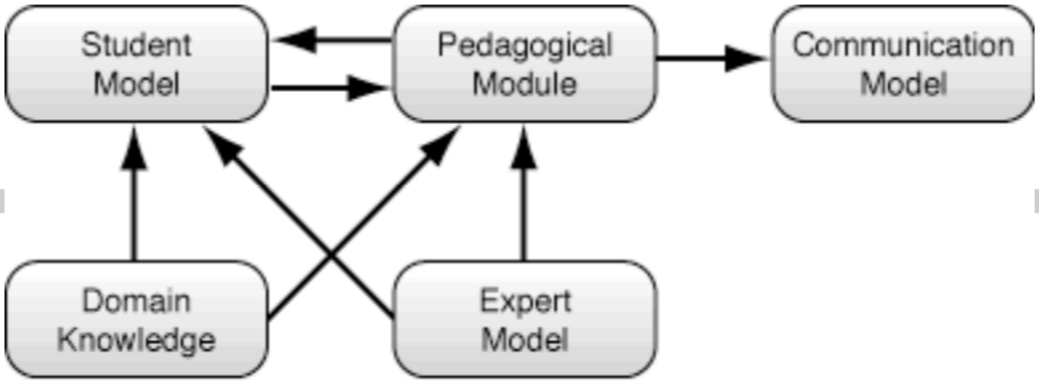
OPEN SOURCE DIALOG SYSTEMS

- Olympus: http://wiki.speech.cs.cmu.edu/olympus (2007)
- InproTK: https://bitbucket.org/inpro/inprotk (2010)
- OpenDial: http://www.opendial-toolkit.net (2012)
- HALEF: http://halef.org (2013)
- Alex: https://ufal.mff.cuni.cz/alex (2014)
- PyDial: http://www.camdial.org/pydial (2017)



Bibauw et al. (2015). Dialogue-based CALL: an overview of existing research

INTELLIGENT TUTORING SYSTEMS



Graesser's 5 steps:

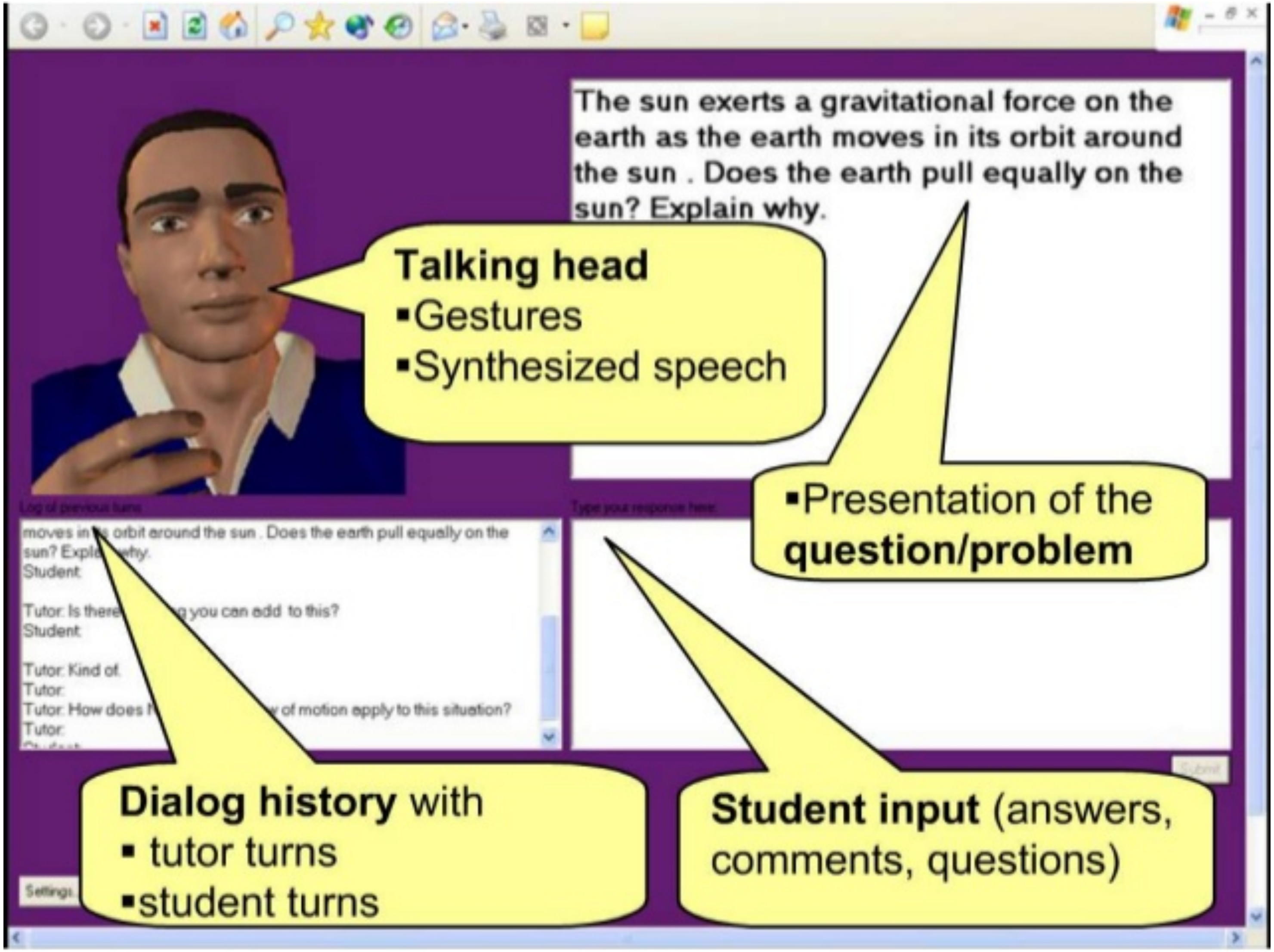
1. Tutor asks question (or presents problem)
2. Learner answers question (or begins to solve problem)
3. Tutor gives short immediate feedback on the quality of the answer (or solution)
4. The tutor and learner collaboratively improve the quality of the answer.
5. The tutor assesses the learner's understanding of the answer

Happens in classroom instruction

Happens in 1-1 tutoring

While many of the early ITSs were dialog-based, they were limited by the NLP technologies at the time.

AUTOTUTOR



The sun exerts a gravitational force on the earth as the earth moves in its orbit around the sun . Does the earth pull equally on the sun? Explain why.

Log of previous turns

moves in its orbit around the sun . Does the earth pull equally on the sun? Expla why.
Student
Tutor: Is there g you can add to this?
Student
Tutor: Kind of.
Tutor:
Tutor: How does t of motion apply to this situation?
Tutor:

Type your response here:

Talking head
- Gestures
- Synthesized speech

- Presentation of the **question/problem**

Dialog history with
- tutor turns
- student turns

Student input (answers, comments, questions)

Developed by Art Graesser and other researchers at the University of Memphis that helps students learn STEAM and ELLA topics through tutorial dialogue in natural language

SDSs for Language Learninig: Designing Conversational Items

10/26/20

Measuring the Power of Learning.™

SDS for Language Learners

- Spoken Dialog Systems (SDS) can help improve foreign language learning and assessment
- SDS enable interactive, conversation-based speaking tasks
  - authentic, situation-based, goal-oriented, dialogic
- benefits for **language learning**:
  - learner can practice conversational English without instructor/tutor present
  - real-time feedback on specific language skills
- benefits for **language assessment**:
  - elicit evidence for aspects of speaking abiliy that can't be measured using traditional, prompt-response items (pragmatics, turn-taking, etc.)

Designing Interactive Speaking Tasks

- Tasks should be naturalistic
- Tasks should be goal-oriented
- Tasks should elicit open-ended speech of variable complexity
- Tasks should be engaging
- Tasks should be robust to off-task responses
- System should adapt to different speakers (L1 backgrounds, proficiency levels, cultural knowledge)
- System should be able to provide feedback about language and task completion

Design Process

- Iterative procedure
  - design initial branching conversation flow based on expected responses
  - collect sample responses from actual users
  - analyze user responses
  - revise conversation flow to handle unmatched responses
  - collect new responses using revised conversation flow
  - ...
- System will never be able to handle every possible response → goal is to make it as robust as possible and handle most likely responses

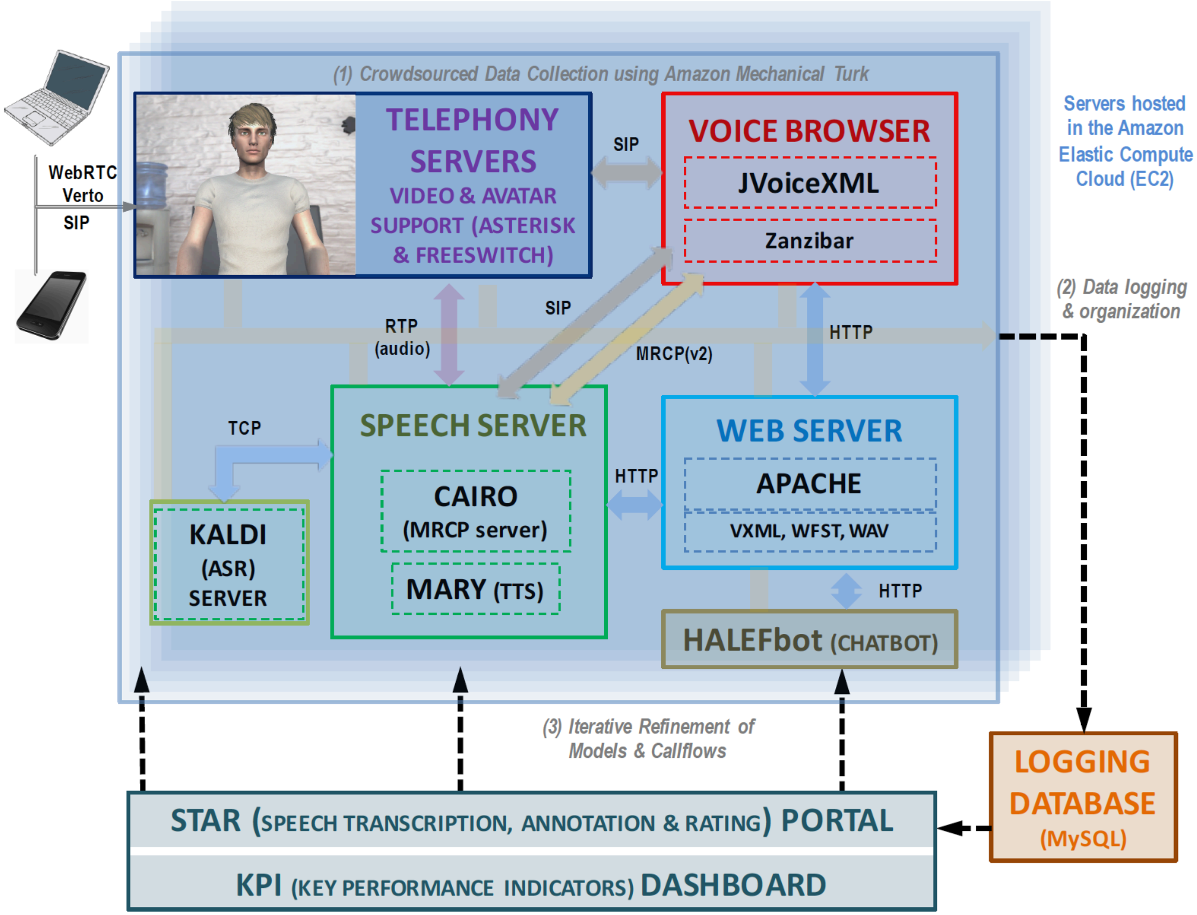
A PROBLEM OF SCALE:

HOW DO WE TEST THE SYSTEM AND COLLECT DATA FROM LARGE NUMBERS OF ACTUAL LEARNERS?

HALEF:
Help Assistant–Language Enabled & Free

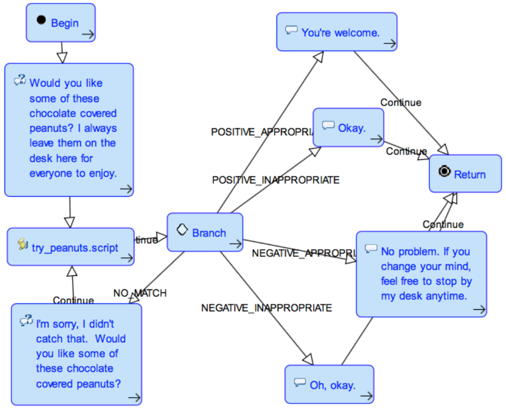
http://halef.org

HALEF is an open-source, standards-compliant, cloud-based and modular dialog system (text, audio or video)

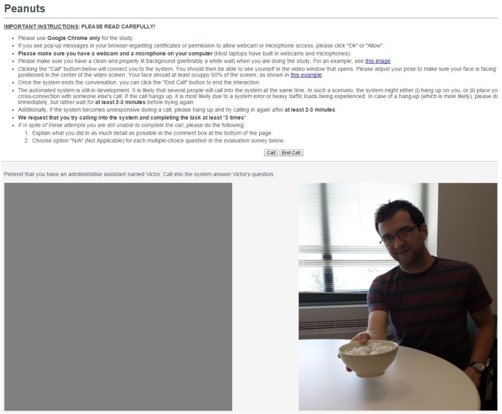

WebRTC
Verto
SIP
(1) Crowdsourced Data Collection using Amazon Mechanical Turk
TELEPHONY SERVERS
VIDEO & AVATAR SUPPORT (ASTERISK & FREESWITCH)
SIP
VOICE BROWSER
JVoiceXML
Zanzibar
Servers hosted in the Amazon Elastic Compute Cloud (EC2)
RTP (audio)
SIP
MRCP(v2)
HTTP
(2) Data logging & organization
TCP
SPEECH SERVER
CAIRO (MRCP server)
MARY (TTS)
HTTP
WEB SERVER
APACHE
VXML, WFST, WAV
KALDI (ASR) SERVER
HTTP
HALEFbot (CHATBOT)
(3) Iterative Refinement of Models & Callflows
LOGGING DATABASE (MySQL)
STAR (SPEECH TRANSCRIPTION, ANNOTATION & RATING) PORTAL
KPI (KEY PERFORMANCE INDICATORS) DASHBOARD

Crowdsourcing techniques allow us to iteratively bootstrap SDSs from scratch & collect data from target populations

TYPICAL DATA COLLECTION PROCESS


Begin
Would you like some of these chocolate covered peanuts? I always leave them on the desk here for everyone to enjoy.
try_peanuts.script
tinue
Continue
NO_MATCH
I'm sorry, I didn't catch that. Would you like some of these chocolate covered peanuts?
Branch
POSITIVE_APPROPRI
POSITIVE_INAPPROPRIATE
NEGATIVE_APPROPRI
NEGATIVE_INAPPROPRIATE
You're welcome.
Okay.
Continue
Continue
Return
Continue
No problem. If you change your mind, feel free to stop by my desk anytime.
Oh, okay.

Design callflows in OpenVXML



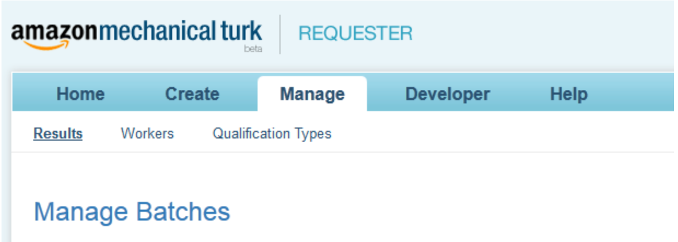
Peanuts

IMPORTANT INSTRUCTIONS: PLEASE READ CAREFULLY!

- Please use **Google Chrome *only*** for the study.
- If you see pop-up messages in your browser regarding certificates or permission to allow webcam or microphone access, please click "Ok" or "Allow".
- **Please make sure you have a webcam and a microphone on your computer** (Most laptops have built in webcams and microphones).
- Please make sure you have a clean and properly lit background (preferably a white wall) when you are doing the study. For an example, see this image.
- Clicking the "Call" buttom below will connect you to the system. You should then be able to see yourself in the video window that opens. Please adjust your pose to make sure your face is facing positioned in the center of the video screen. Your face should at least ocuppy 60% of the screen, as shown in this example.
- Once the system ends the conversation, you can click the "End Call" button to end the interaction.
- The automated system is still in development. It is likely that several people will call into the system at the same time. In such a scenario, the system might either (i) hang up on you, or (ii) place yo cross-connection with someone else's call. If the call hangs up, it is most likely due to a system error or heavy traffic loads being experienced. In case of a hang-up (which is more likely), please do immediately, but rather wait for **at least 2-3 minutes** before trying again.
- Additionally, if the system becomes unresponsive during a call, please hang up and try calling in again after **at least 2-3 minutes**.
- **We request that you try calling into the system and completing the task at least "3 times"**.
- *If in spite of these attempts you are still unable to complete the call*, please do the following:
  1. Explain what you did in as much detail as possible in the comment box at the bottom of the page.
  2. Choose option "N/A" (Not Applicable) for each multiple-choice question in the evaluation survey below.

Pretend that you have an administrative assistant named Victor. Call into the system answer Victor's question.

Design webpage interface



Create task on crowdsourcing platform

Proof of task completion

Making Requests

- Communicative function: making requests in a workplace environment
- Target language: pragmatically appropriate phrases for indirect requests
- Task: learner plays the role of an office worker having a conversation with their boss and is instructed to make two requests

Task Directions

Imagine that you are calling your boss, Lisa Green. Your goals are to:

1. get her to agree to have a meeting with you and
2. ask her to review your presentation slides before the meeting.

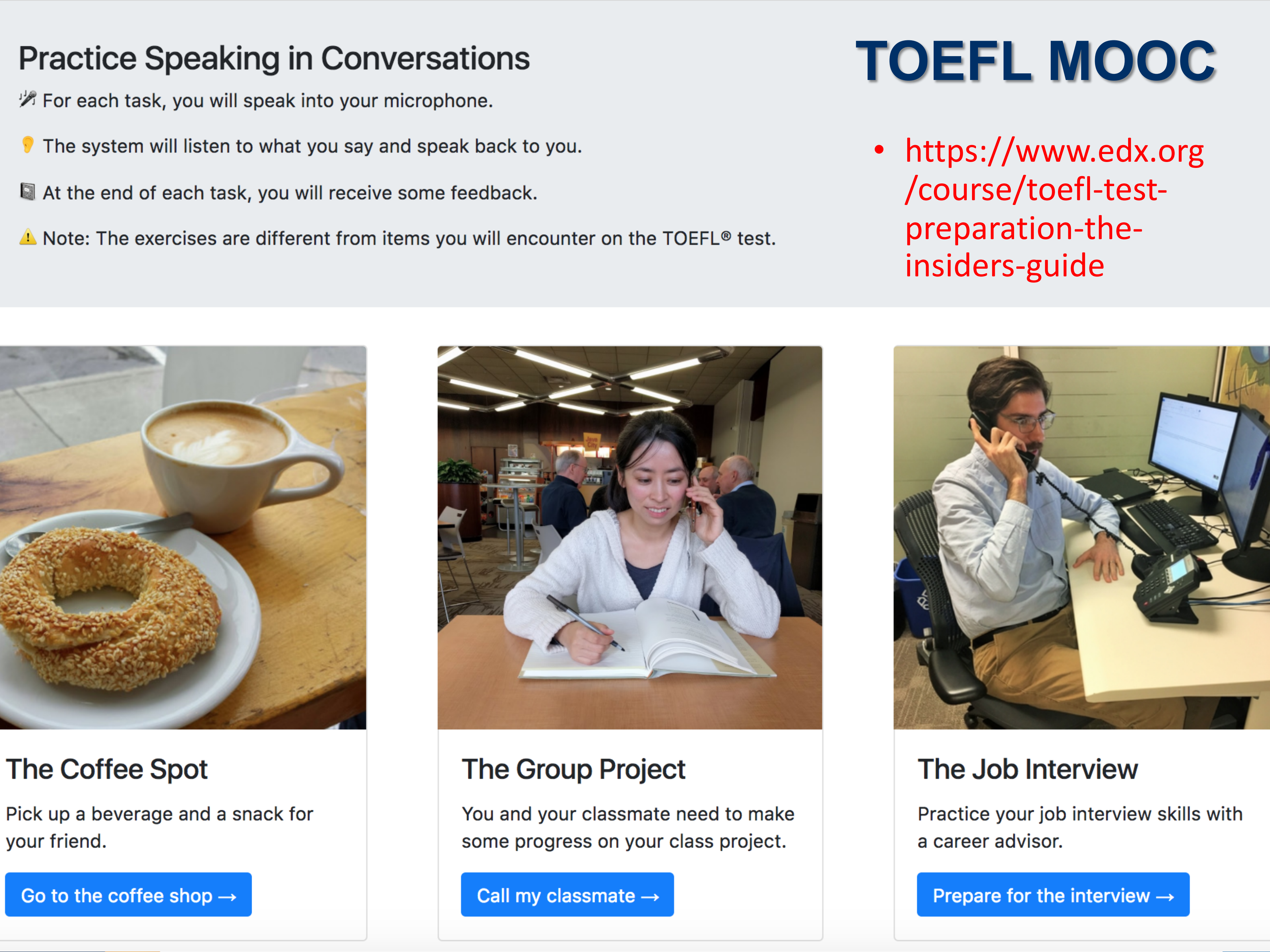
Your schedule is free for the rest of the week, so any time proposed by Lisa will work for you.

You:

Lisa Green:

TOEFL MOOC

- https://www.edx.org/course/toefl-test-preparation-the-insiders-guide



Practice Speaking in Conversations

🎤 For each task, you will speak into your microphone.

💡 The system will listen to what you say and speak back to you.

📓 At the end of each task, you will receive some feedback.

⚠️ Note: The exercises are different from items you will encounter on the TOEFL® test.

The Coffee Spot

Pick up a beverage and a snack for your friend.

The Group Project

You and your classmate need to make some progress on your class project.

The Job Interview

Practice your job interview skills with a career advisor.

FEEDBACK & SCORING

SCORING INTERACTIONAL COMPETENCE

- *Big Picture Goal*: developing intelligent dialog agents for instruction, learning and assessment of learners of English
  - Conversational proficiency is a crucial skill in today's workplace, but is not an easy construct to define (Young, 2011; Doehler and Pochon-Berger, 2015).
  - Dialog systems offer one way of automating this need at scale.

- *Study Motivation*: no prior study has performed a comprehensive examination of the automated scoring of whole dialog responses based primarily on text features, specifically for interaction aspects thereof.

- *Key contributions*:
  1. developing a comprehensive rubric specifically tailored to conversational dialog
  2. triple-scoring a selection of dialog data based on this rubric
  3. examining the performance of two methods for automated scoring– the first based on interpretable feature engineering and the second based on deep model engineering.

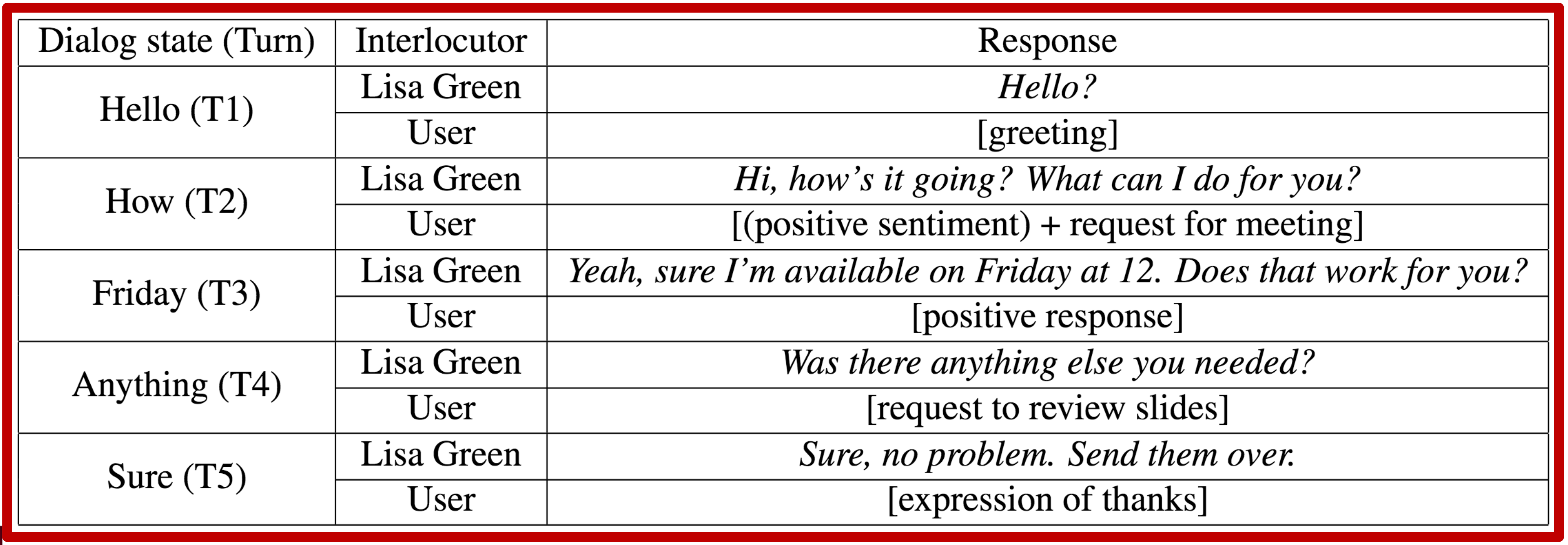
EXPERIMENTAL SETUP

- 2288 conversations of human-machine dialog (Boss – Employee interaction)

- Deployed on Amazon Mechanical Turk using the HALEF open source dialog system

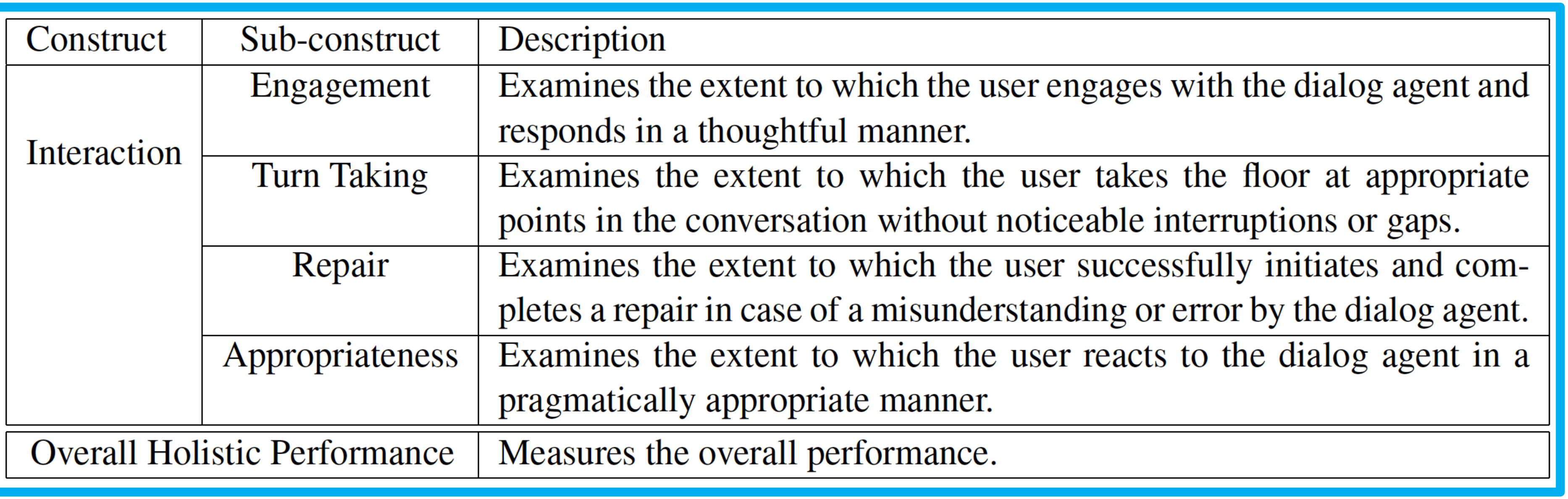
- Human Scoring: Each of the 2288 dialog responses triple scored (using a randomized design) by 3 of 8 human raters on a custom-designed rubric.

- Machine Scoring: 10-fold cross-validation experiments
    - Input: full dialog (all turns) → Output: Scores.
    - Feature engineering and e2e deep learning model engineering explored



Dialog state (Turn)	Interlocutor	Response
Hello (T1)	Lisa Green	*Hello?*
	User	[greeting]
How (T2)	Lisa Green	*Hi, how's it going? What can I do for you?*
	User	[(positive sentiment) + request for meeting]
Friday (T3)	Lisa Green	*Yeah, sure I'm available on Friday at 12. Does that work for you?*
	User	[positive response]
Anything (T4)	Lisa Green	*Was there anything else you needed?*
	User	[request to review slides]
Sure (T5)	Lisa Green	*Sure, no problem. Send them over.*
	User	[expression of thanks]

SCORING RUBRIC

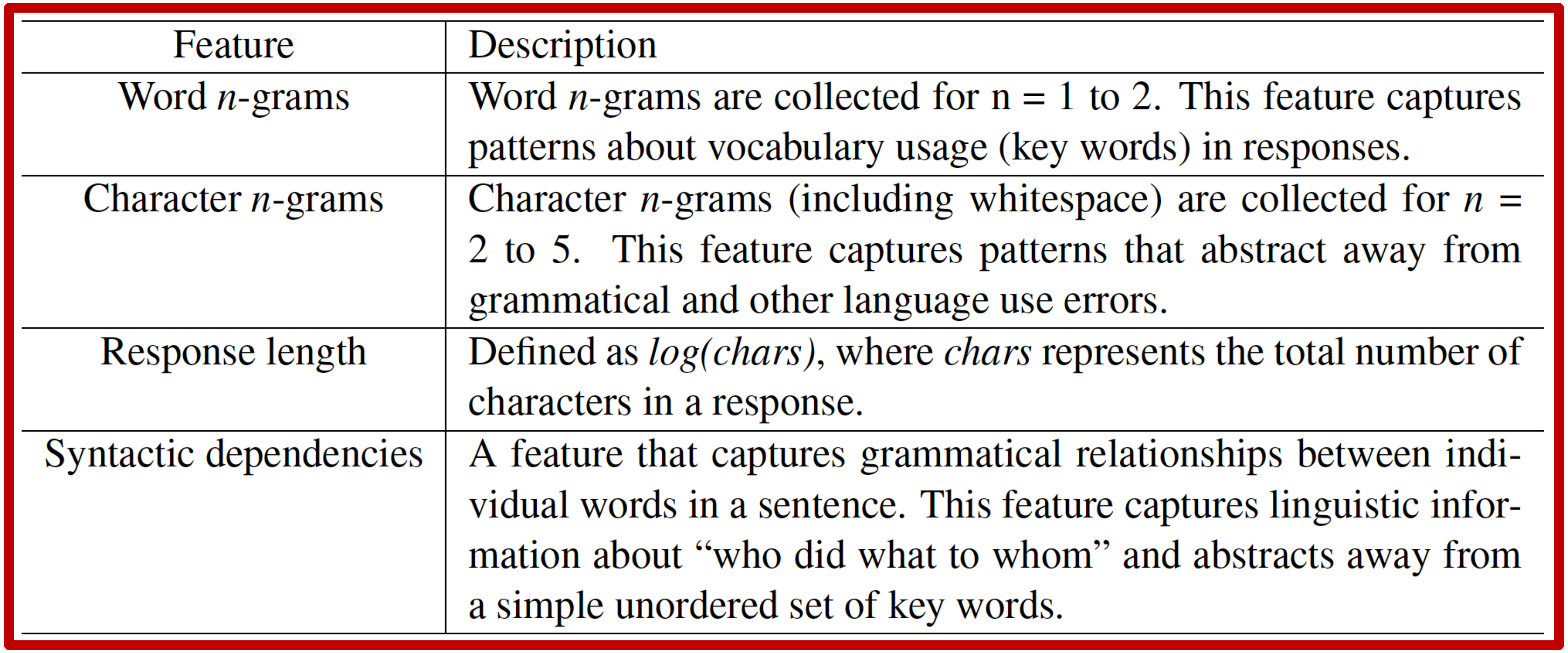


Construct	Sub-construct	Description
Interaction	Engagement	Examines the extent to which the user engages with the dialog agent and responds in a thoughtful manner.
	Turn Taking	Examines the extent to which the user takes the floor at appropriate points in the conversation without noticeable interruptions or gaps.
	Repair	Examines the extent to which the user successfully initiates and completes a repair in case of a misunderstanding or error by the dialog agent.
	Appropriateness	Examines the extent to which the user reacts to the dialog agent in a pragmatically appropriate manner.
Overall Holistic Performance		Measures the overall performance.

Human scoring rubric for interaction aspects of conversational proficiency. Scores are assigned on a Likert scale from 1-4 ranging from low to high proficiency. A score of 0 is assigned when there were issues with audio quality or system malfunction or off-topic or empty responses.

Also see Ramanarayanan (2020). Design and Development of a Human-Machine Dialog Corpus for the Automated Assessment of Conversational English Proficiency

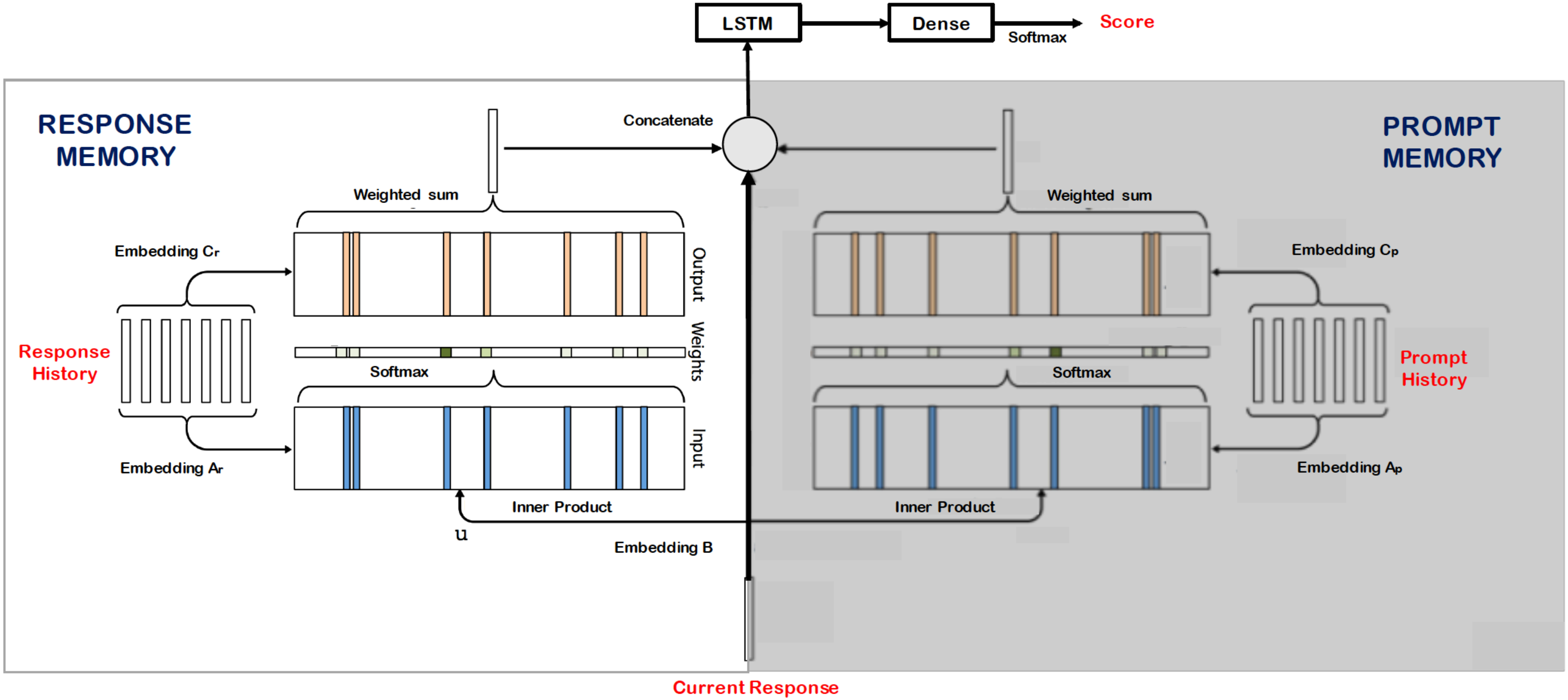
FEATURE ENGINEERING



Feature	Description
Word *n*-grams	Word *n*-grams are collected for n = 1 to 2. This feature captures patterns about vocabulary usage (key words) in responses.
Character *n*-grams	Character *n*-grams (including whitespace) are collected for *n* = 2 to 5. This feature captures patterns that abstract away from grammatical and other language use errors.
Response length	Defined as *log(chars)*, where *chars* represents the total number of characters in a response.
Syntactic dependencies	A feature that captures grammatical relationships between individual words in a sentence. This feature captures linguistic information about "who did what to whom" and abstracts away from a simple unordered set of key words.

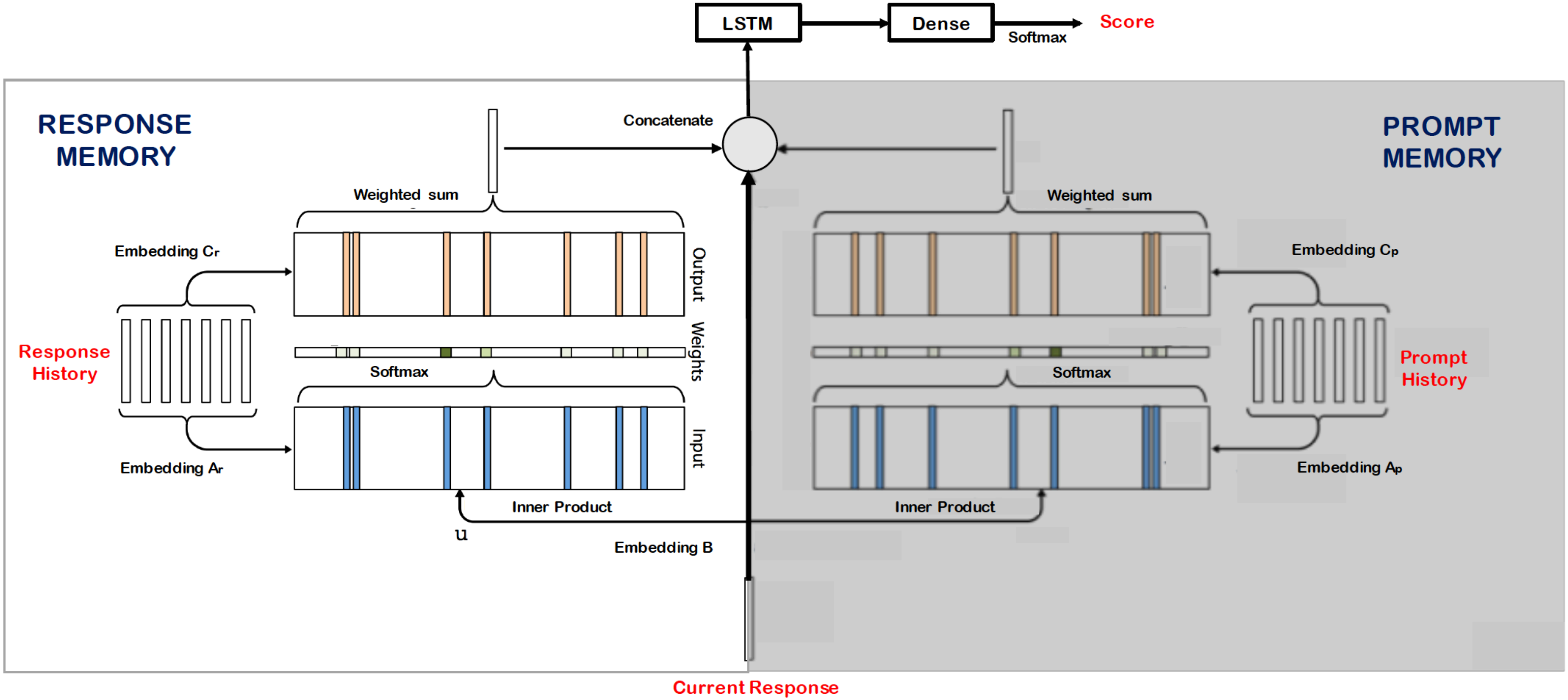
c-rater NLP features used for machine scoring (Madnani et al., 2017)

MEMORY NETWORKS



The MemN2N architecture Sukhbaatar et al. (2015) learns a mapping between an output score and an input tuple consisting of the current response, the response history and the prompt history.

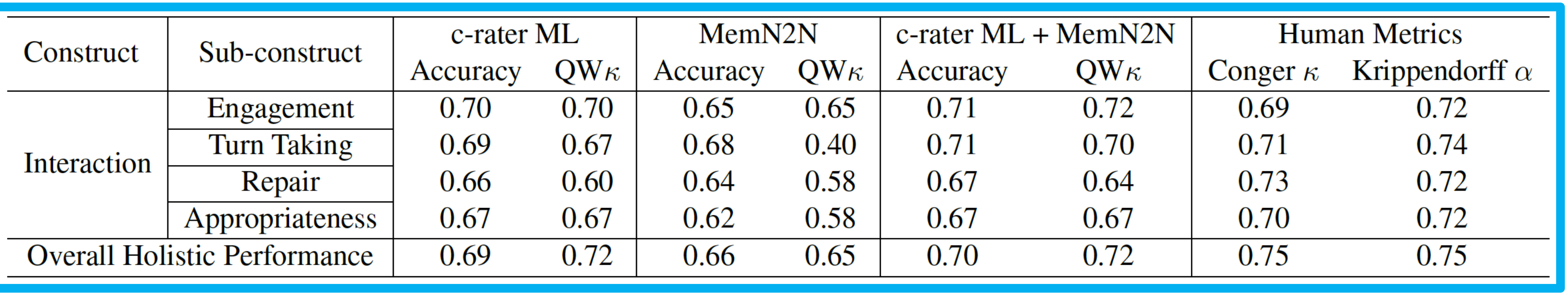
MEMORY NETWORKS



MemN2N training details:

- Optimized a cross-entropy-based objective function
- Tuned hyperparameters (# of layers, neurons per layer, dropout) using *hyperas*.
- Experimented with 1, 2 and 3 memory hops and found 2 to be optimal.
- Pretrained embeddings worked better than random for prompt history encoding

RESULTS



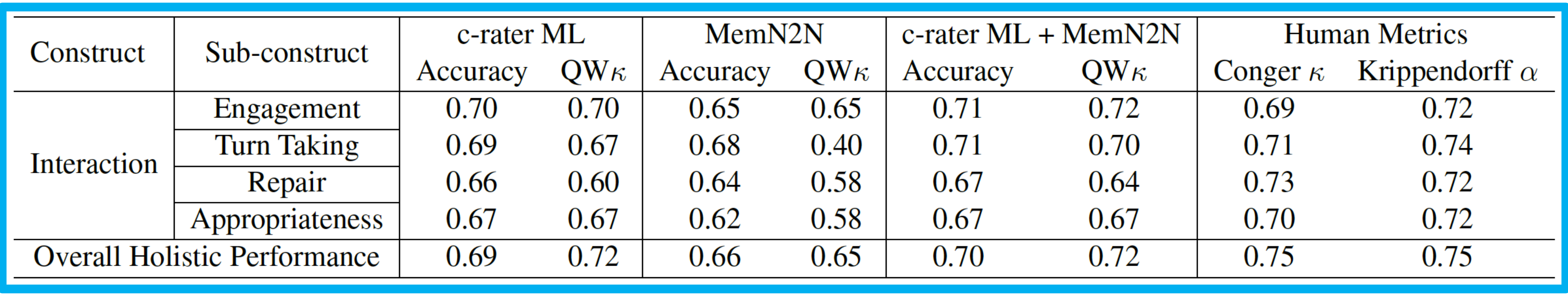
Construct	Sub-construct	c-rater ML		MemN2N		c-rater ML + MemN2N		Human Metrics	
		Accuracy	QWκ	Accuracy	QWκ	Accuracy	QWκ	Conger κ	Krippendorff α
Interaction	Engagement	0.70	0.70	0.65	0.65	0.71	0.72	0.69	0.72
	Turn Taking	0.69	0.67	0.68	0.40	0.71	0.70	0.71	0.74
	Repair	0.66	0.60	0.64	0.58	0.67	0.64	0.73	0.72
	Appropriateness	0.67	0.67	0.62	0.58	0.67	0.67	0.70	0.72
Overall Holistic Performance		0.69	0.72	0.66	0.65	0.70	0.72	0.75	0.75

Moderate to high agreement among human raters.

Fusing the MemN2N with the c-rater ML system leads to a small but significant improvement over either of the systems alone.

A combination of n-gram, length, syntactic dependency and memory-based attention over embedding representations of words over the entire dialog are useful in capturing at least some aspects of these sub-constructs of interaction.

RESULTS



Construct	Sub-construct	c-rater ML		MemN2N		c-rater ML + MemN2N		Human Metrics	
		Accuracy	QWκ	Accuracy	QWκ	Accuracy	QWκ	Conger κ	Krippendorff α
Interaction	Engagement	0.70	0.70	0.65	0.65	0.71	0.72	0.69	0.72
	Turn Taking	0.69	0.67	0.68	0.40	0.71	0.70	0.71	0.74
	Repair	0.66	0.60	0.64	0.58	0.67	0.64	0.73	0.72
	Appropriateness	0.67	0.67	0.62	0.58	0.67	0.67	0.70	0.72
Overall Holistic Performance		0.69	0.72	0.66	0.65	0.70	0.72	0.75	0.75

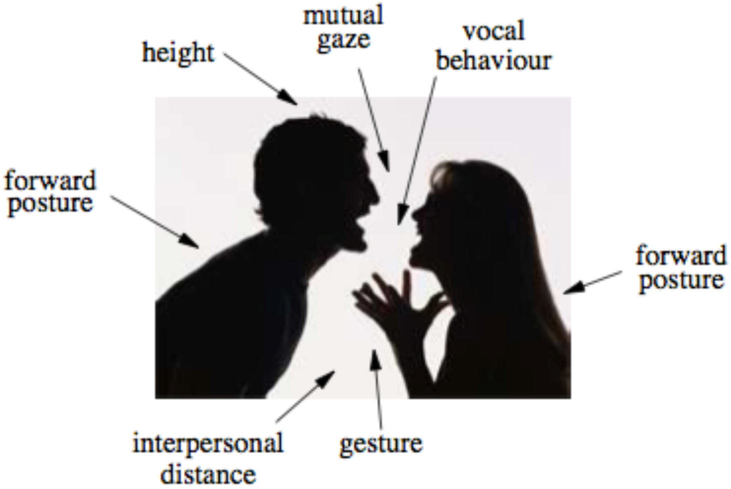
Repair and appropriateness, and even turn taking to a lesser extent are related to proficiency in language use, and hence it makes sense that features such as n-grams and syntactic dependencies might be somewhat useful.

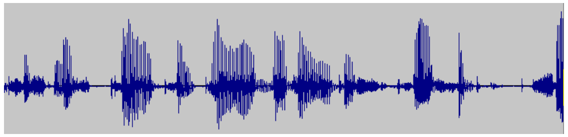
However, some of the results might also be explained by some of our examined features being highly correlated with more interpretable/relevant features.

Future: consider information from audio or visual channels; useful in predicting properties related to interaction (engagement, for instance).

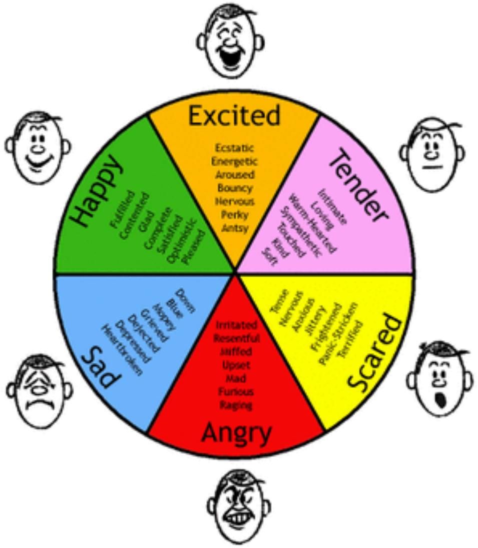
MULTIMODAL ANALYSIS

As the name suggests, we use data from multiple modalities to inform automatic learning and assessment problems

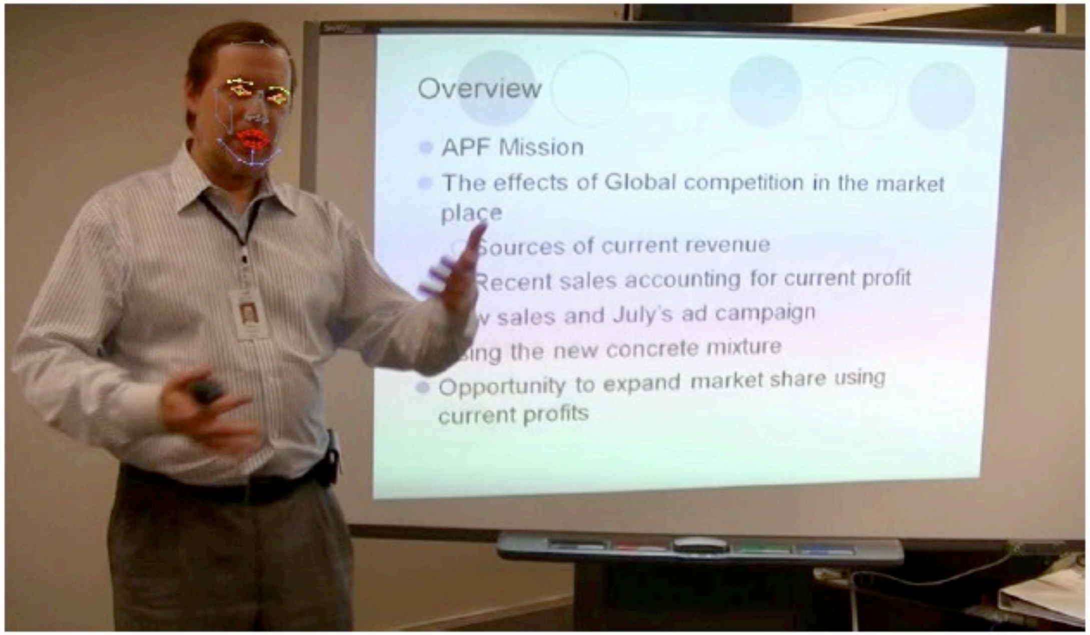


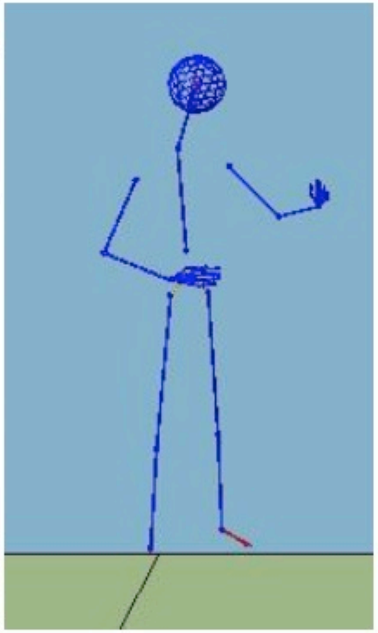


TEXT



EXAMPLE APPLICATION AT ETS:
MULTIMODAL PRESENTATION SCORING


Overview
APF Mission
The effects of Global competition in the market place
Sources of current revenue
Recent sales accounting for current profit
w sales and July's ad campaign
ing the new concrete mixture
Opportunity to expand market share using current profits



IDEA: Use speech, visual and body movement (Microsoft Kinect) data to automatically predict different scores related to presentation proficiency

PRESENTATION TASKS

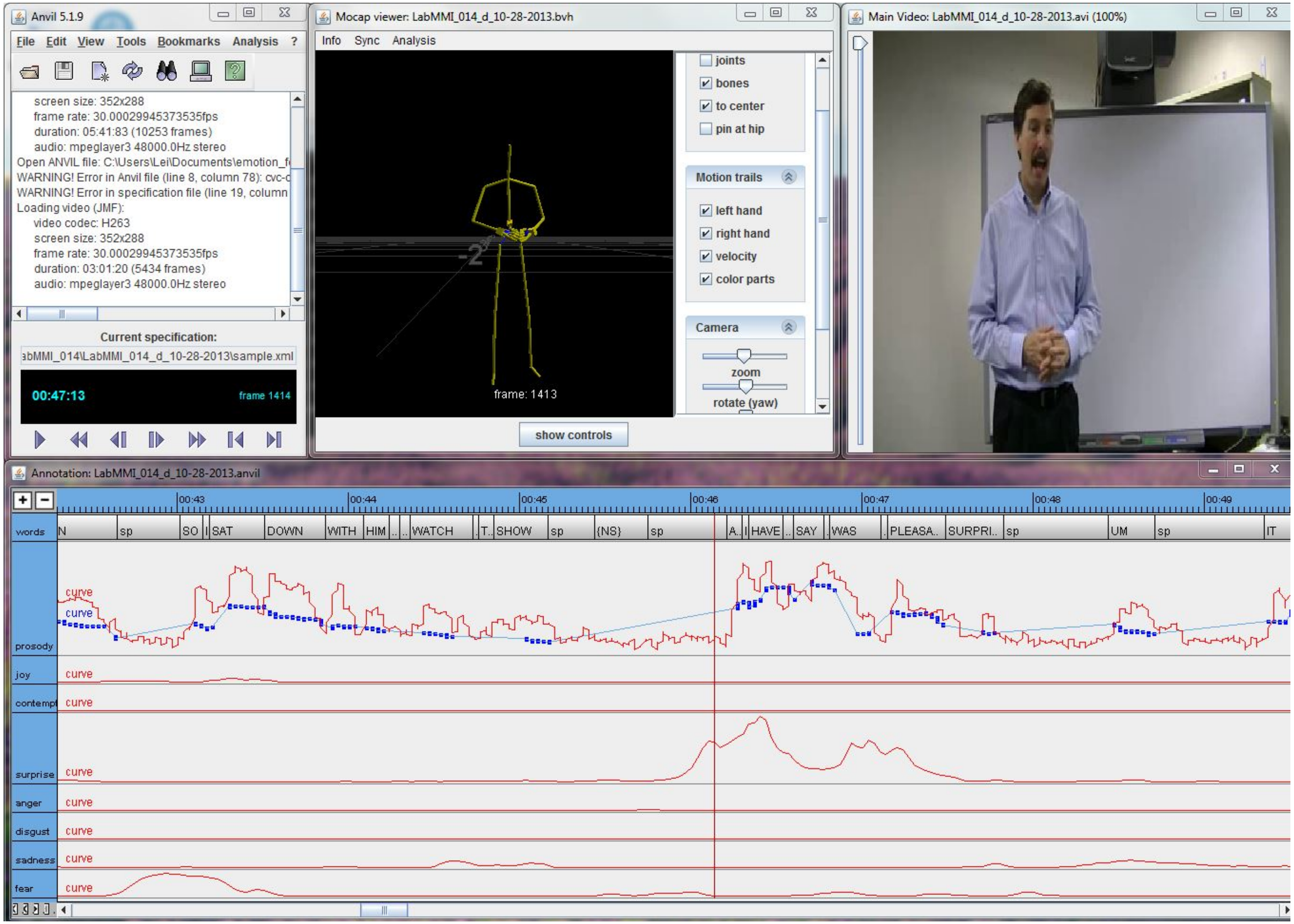
1. Speaker given a slide deck and ~10 minutes for preparation to :
   a) present a financial report
   b) teach a topic targeted at middle school students

2. Impromtu speeches on:
   a) a movie you did not like but have to recommend nonetheless
   b) the benefits of a place that is typically inconvenient to live in

MULTIMODAL DATA COLLECTION

- 14 speakers – 6 males, 8 females

- Multiple data streams synchronized
  - Kinect
  - Speech
  - Face
  - Emotion

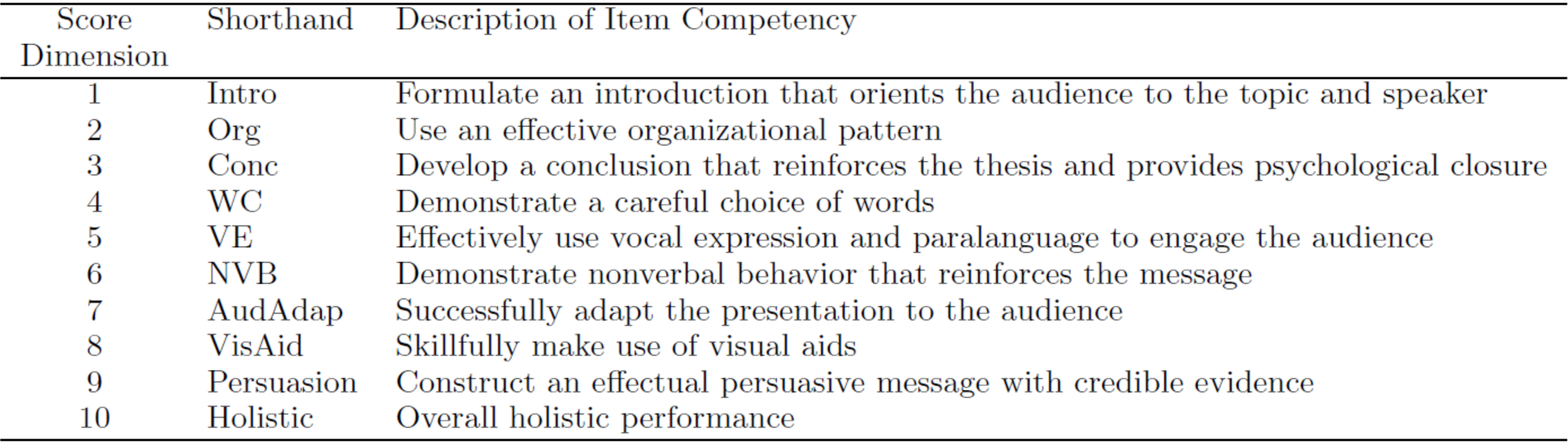


```
    screen size: 352x288
    frame rate: 30.00029945373535fps
    duration: 05:41:83 (10253 frames)
    audio: mpeglayer3 48000.0Hz stereo
Open ANVIL file: C:\Users\Lei\Documents\emotion_f
WARNING! Error in Anvil file (line 8, column 78): cvc-c
WARNING! Error in specification file (line 19, column
Loading video (JMF):
    video codec: H263
    screen size: 352x288
    frame rate: 30.00029945373535fps
    duration: 03:01:20 (5434 frames)
    audio: mpeglayer3 48000.0Hz stereo
```

Current specification: abMMI_014\LabMMI_014_d_10-28-2013\sample.xml

SCORING RUBRIC

Atleast 2 human raters scored all 56 (4x14) presentation videos on these 10 aspects of the Public Speaking Competence Rubric (PSCR) on a 5-point Likert scale (0-4).

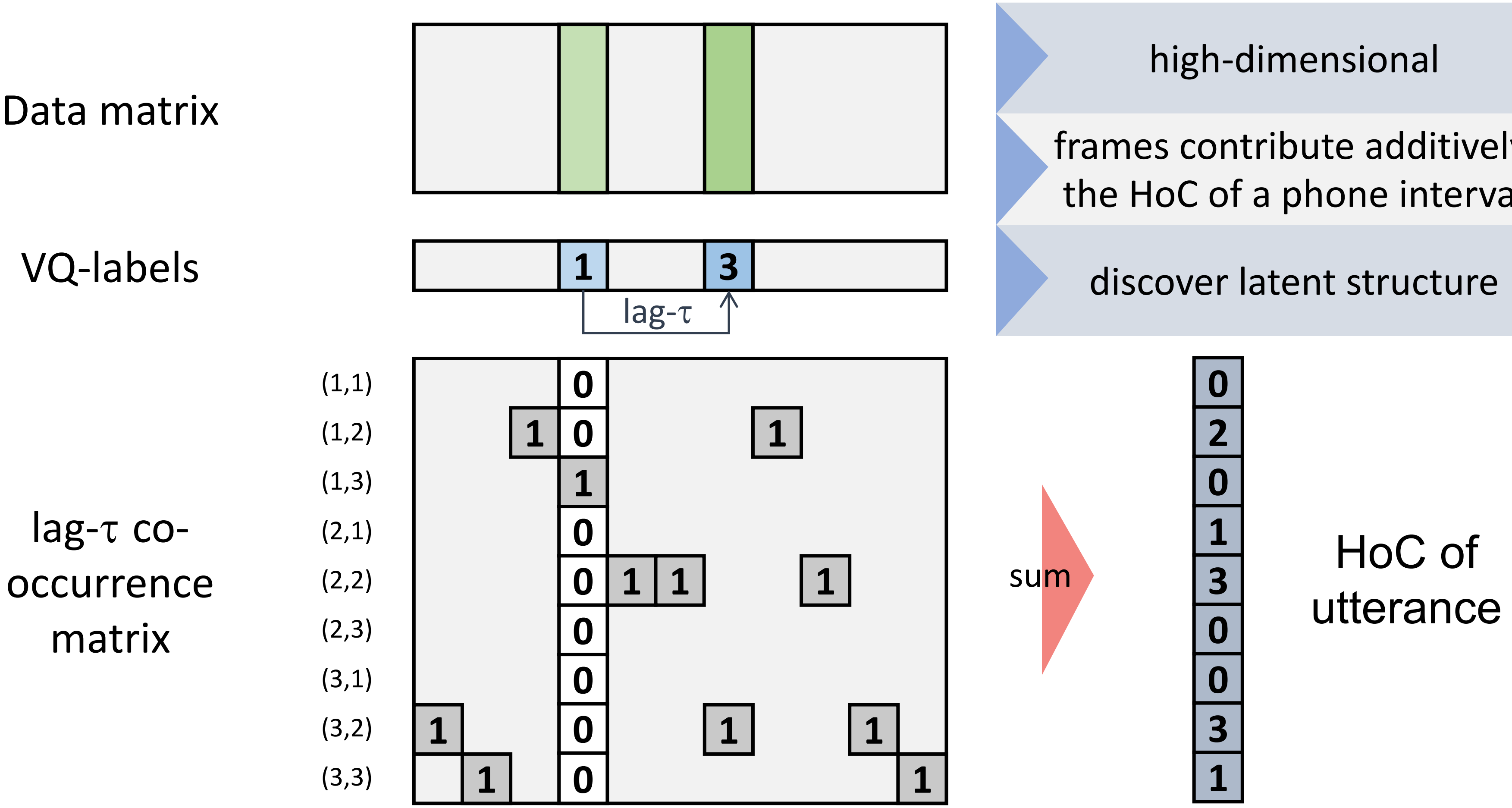


Score Dimension	Shorthand	Description of Item Competency
1	Intro	Formulate an introduction that orients the audience to the topic and speaker
2	Org	Use an effective organizational pattern
3	Conc	Develop a conclusion that reinforces the thesis and provides psychological closure
4	WC	Demonstrate a careful choice of words
5	VE	Effectively use vocal expression and paralanguage to engage the audience
6	NVB	Demonstrate nonverbal behavior that reinforces the message
7	AudAdap	Successfully adapt the presentation to the audience
8	VisAid	Skillfully make use of visual aids
9	Persuasion	Construct an effectual persuasive message with credible evidence
10	Holistic	Overall holistic performance

If the two raters did not agree, then a third rater was brought in, and the final score was the average of all three.

L. M. Schreiber, G. D. Paul, and L. R. Shibley. The development and test of the public speaking competence rubric. Communication Education, 61(3):205{233, 2012.

HISTOGRAMS OF COOCCURRENCES (HoC)



- high-dimensional
- frames contribute additively to the HoC of a phone interval
- discover latent structure

HAC features compactly encapsulate **spatiotemporal** information in each phone interval!

SPEECH PROFICIENCY FEATURES

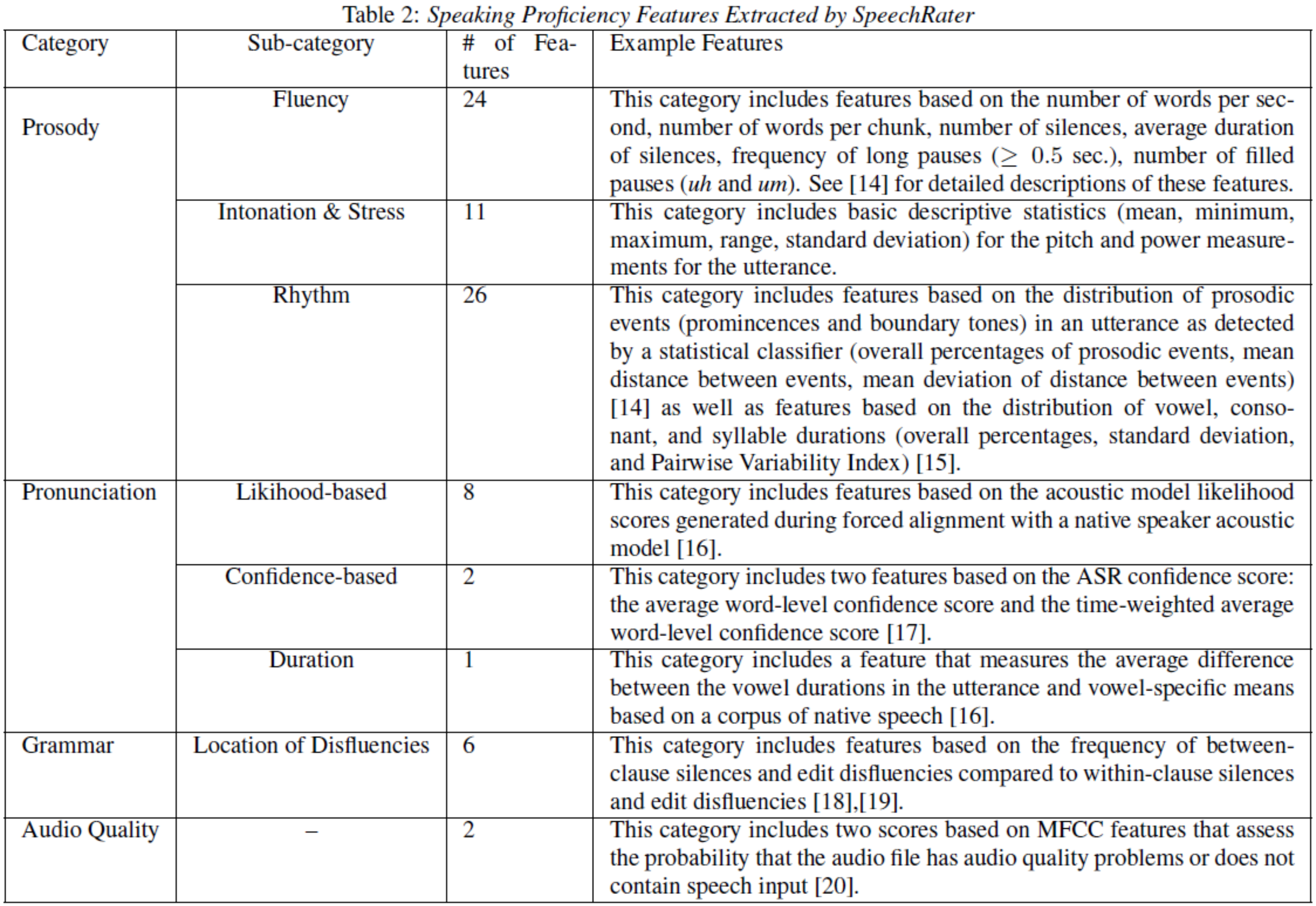
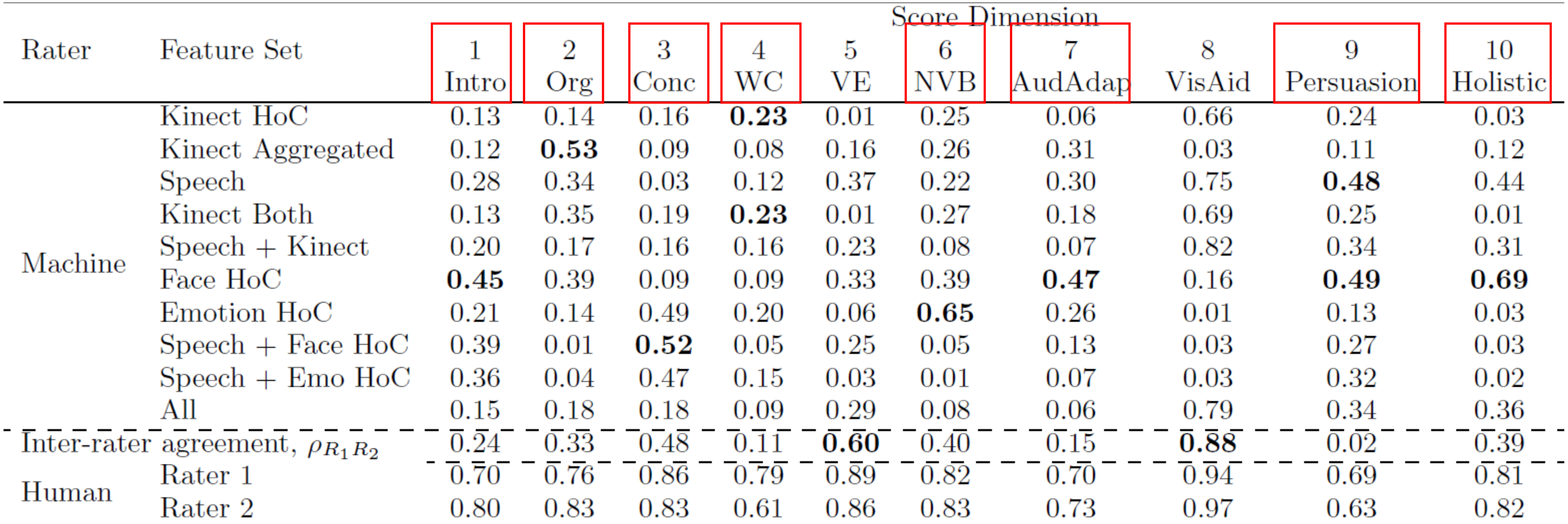


Table 2: *Speaking Proficiency Features Extracted by SpeechRater*

Category	Sub-category	# of Features	Example Features
Prosody	Fluency	24	This category includes features based on the number of words per second, number of words per chunk, number of silences, average duration of silences, frequency of long pauses (≥ 0.5 sec.), number of filled pauses (*uh* and *um*). See [14] for detailed descriptions of these features.
	Intonation & Stress	11	This category includes basic descriptive statistics (mean, minimum, maximum, range, standard deviation) for the pitch and power measurements for the utterance.
	Rhythm	26	This category includes features based on the distribution of prosodic events (promincences and boundary tones) in an utterance as detected by a statistical classifier (overall percentages of prosodic events, mean distance between events, mean deviation of distance between events) [14] as well as features based on the distribution of vowel, consonant, and syllable durations (overall percentages, standard deviation, and Pairwise Variability Index) [15].
Pronunciation	Likihood-based	8	This category includes features based on the acoustic model likelihood scores generated during forced alignment with a native speaker acoustic model [16].
	Confidence-based	2	This category includes two features based on the ASR confidence score: the average word-level confidence score and the time-weighted average word-level confidence score [17].
	Duration	1	This category includes a feature that measures the average difference between the vowel durations in the utterance and vowel-specific means based on a corpus of native speech [16].
Grammar	Location of Disfluencies	6	This category includes features based on the frequency of between-clause silences and edit disfluencies compared to within-clause silences and edit disfluencies [18],[19].
Audio Quality	–	2	This category includes two scores based on MFCC features that assess the probability that the audio file has audio quality problems or does not contain speech input [20].

RESULTS



Rater	Feature Set	Score Dimension									
		1	2	3	4	5	6	7	8	9	10
		Intro	Org	Conc	WC	VE	NVB	AudAdap	VisAid	Persuasion	Holistic
Machine	Kinect HoC	0.13	0.14	0.16	**0.23**	0.01	0.25	0.06	0.66	0.24	0.03
	Kinect Aggregated	0.12	**0.53**	0.09	0.08	0.16	0.26	0.31	0.03	0.11	0.12
	Speech	0.28	0.34	0.03	0.12	0.37	0.22	0.30	0.75	**0.48**	0.44
	Kinect Both	0.13	0.35	0.19	**0.23**	0.01	0.27	0.18	0.69	0.25	0.01
	Speech + Kinect	0.20	0.17	0.16	0.16	0.23	0.08	0.07	0.82	0.34	0.31
	Face HoC	**0.45**	0.39	0.09	0.09	0.33	0.39	**0.47**	0.16	**0.49**	**0.69**
	Emotion HoC	0.21	0.14	0.49	0.20	0.06	**0.65**	0.26	0.01	0.13	0.03
	Speech + Face HoC	0.39	0.01	**0.52**	0.05	0.25	0.05	0.13	0.03	0.27	0.03
	Speech + Emo HoC	0.36	0.04	0.47	0.15	0.03	0.01	0.07	0.03	0.32	0.02
	All	0.15	0.18	0.18	0.09	0.29	0.08	0.06	0.79	0.34	0.36
Inter-rater agreement, $\rho_{R_1 R_2}$		0.24	0.33	0.48	0.11	**0.60**	0.40	0.15	**0.88**	0.02	0.39
Human	Rater 1	0.70	0.76	0.86	0.79	0.89	0.82	0.70	0.94	0.69	0.81
	Rater 2	0.80	0.83	0.83	0.61	0.86	0.83	0.73	0.97	0.63	0.82

The machine predicted score had a higher correlation with the final score than the agreement between the first two human raters in 8/10 cases!

While it is clear why some features performed well (emotion, face features for NVB and Holistic score), others are less interpretable (→ future work).

WRAPPING UP…

KEY TAKEAWAYS

1. Several factors to consider for L2 speech assessment
   - Acoustics, L1, Culture, Task Design, Score Type, Demographic

2. Typical steps of automated scoring
   - Recorded response → Speech recognition → Feature computation → Filtering → Scoring Model

3. Careful design of rubrics is important!
   - Example dimensions: Delivery, Language Use, Topic Development, Interaction/Kinesics (dialog)

4. Reliability, Validity and Fairness are crucial and often overlooked considerations for large-scale scoring systems

5. As we move to more interactive technologies for learning and assessment, dialog/multimodal scoring become increasingly important

References

- Bernstein, J., Cohen, M., Murveit, H., Rtischev, D., & Weintraub, M. (1990). Automatic Evaluation and Training in English Pronunciation. In Proceedings of ICSLP 90 (pp. 1185–1188).
- Bibauw et al. (2015). Dialogue-based CALL: an overview of existing research
- Chen, L., Zechner, K., Yoon, S.-Y., Evanini, K., Wang, X., Loukina, A., Tao, J., Davis, L., Lee, C.M., Ma, M., Mundkowsky, R., Lu, C., Leong, C.W., & Gyawali, B. (2018). *Automated scoring of nonnative speech using the SpeechRater v. 5.0 engine* (Research Report No. RR-18-10). Princeton, NJ: Educational Testing Service.
- Cucchiarini, C., Strik, H., and 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):2862-73. July 2002.
- Gu, L., Davis, L., Tao, J., & Zechner, K. (2020). Using spoken language technology for generating feedback to prepare for the TOEFL iBT® test: a user perception study. *Assessment in Education: Principles, Policy & Practice.*
- Haberman, S. J. (2008). When can subscores have value? *Journal of Educational and Behavioral Statistics*, 33, 204–229.

References

- Lehman, B., Gu, L., Zhao, J., Tsuprun, E., Kurzum, C., Schiano, M., Liu, Y., & Jackson, G.T. (2020). Use of adaptive feedback in an app for English language spontaneous speech. In *Proceedings of AIED 2020*, 309-320.
- Nitin Madnani, Anastassia Loukina, and Aoife Cahill.2017. A large scale quantitative exploration of mod-eling strategies for content scoring. In Proceedings of the 12th Workshop on Innovative Use of NLP forBuilding Educational Applications, pages 457–467.
- Yao Qian , Patrick Lange , Keelan Evanini , Robert Pugh , Rutuja Ubale, Matthew Mulholland , Xinhao Wang (2019). Neural Approaches to Automated Speech Scoring of Monologue and Dialogue Responses.: ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 12-17 May 2019
- Yao Qian, Rutuja Ubale, Patrick Lange, Keelan Evanini, Vikram Ramanarayanan and Frank K. Soong (2019). Spoken Language Understanding of Human-Machine Conversations for Language Learning Applications, in: Journal of Signal Processing Systems.
- Vikram Ramanarayanan (2020). Design and Development of a Human-Machine Dialog Corpus for the Automated Assessment of Conversational English Proficiency, in proceedings of: Interspeech 2020, Virtual Conference, Oct 2020.
- Vikram Ramanarayanan, Matt Mulholland and Debanjan Ghosh (2020). Exploring Recurrent, Memory and Attention Based Architectures for Scoring Interactional Aspects of Human-Machine Text Dialog.

References

- Vikram Ramanarayanan, David Suendermann-Oeft, Patrick Lange, Alexei V. Ivanov, Keelan Evanini, Zhou Yu, Eugene Tsuprun, and Yao Qian (2016), Bootstrapping Development of a Cloud-Based Spoken Dialog System in the Educational Domain From Scratch Using Crowdsourced Data, in: ETS Research Report Series, Wiley. doi: 10.1002/ets2.12105.
- Vikram Ramanarayanan, Chee Wee Leong, Lei Chen, Gary Feng and David Suendermann-Oeft (2015). Evaluating speech, face, emotion and body movement time-series features for automated multimodal presentation scoring, in proceedings of: International Conference on Multimodal Interaction (ICMI 2015), Seattle, WA, Nov 2015.
- Vikram Ramanarayanan, Keelan Evanini and Eugene Tsuprun (2019), Beyond monologues: Automated processing of conversational speech, in: Automated Speaking Assessment: Using Language Technologies to Score Spontaneous Speech., K. Zechner and K. Evanini, Eds., London: Routledge - Taylor and Francis.
- Sinharay, S., Puhan, G., & Haberman, S. J. (2011). An NCME instructional module on subscores. *Educational Measurement: Issues and Practice*, 30(3), 29-40.
- Wang, X., Zechner, K., and Hamill, C. O. (2020). Targeted Content Feedback in Spoken Language Learning and Assessment. InterSpeech-2020, Shanghai, China, October.

References

- Yu, Z., Ramanarayanan, V., Suendermann-Oeft, D., Wang, X., Zechner, K., Chen, L., Tao, J., & Qian, Y. (2015). Using bidirectional LSTM recurrent neural networks to learn high-level abstractions of sequential features for automated scoring of non-native spontaneous speech. Proceedings of Automatic Speech Recognition and Understanding workshop (ASRU-2015), Scotsdale, AZ, December.

- 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.

- Zechner, K. & Evanini, K. (Eds.) (2019). Automated Speaking Assessment: Using Language technologies to Score Spontaneous Speech. In J. Norris, J.E. Purpura, S.J. Ross, & X. Xi (Eds.): Innovations in Language Learning and Assessment at ETS (Vol. 3). New York, NY: Routledge.

QUESTIONS?

vramanarayanan@ets.org
kzechner@ets.org
kevanini@ets.org