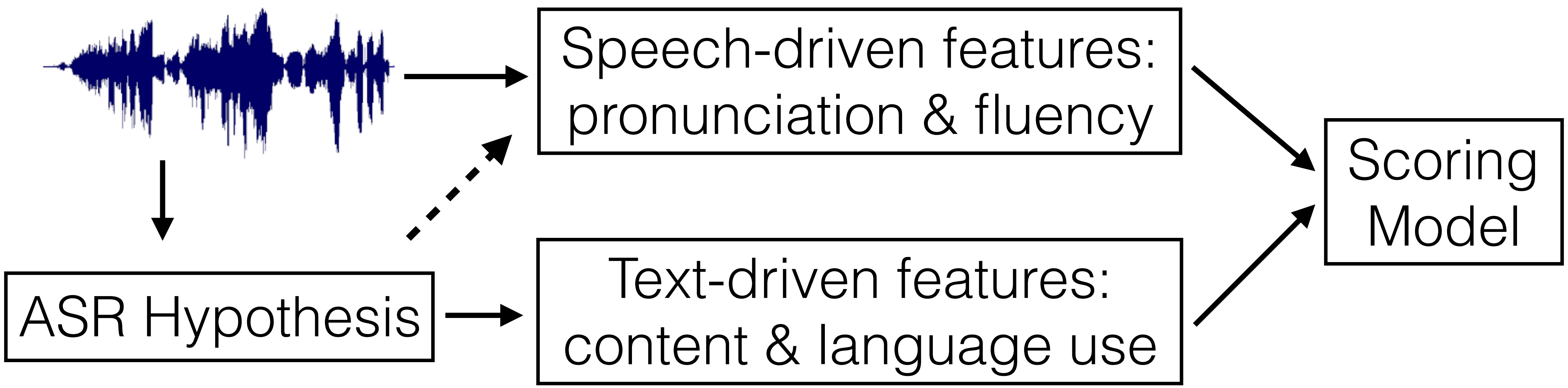
Speech- and Text-Driven Features for Automated Scoring of English Speaking Tasks

Anastassia Loukina
Nitin Madnani
Aoife Cahill

Copyright © 2017 by Educational Testing Service. All rights reserved. ETS and the ETS logo are registered trademarks of Educational Testing Service (ETS). MEASURING THE POWER OF LEARNING is a trademark of ETS. 30141

Automatically Scoring Spoken Responses

- Goal: assess test-taker's language proficiency
- "*Did the test-taker produce a coherent, intelligible response that addresses the question?*"



Copyright © 2017 by Educational Testing Service. All rights reserved. ETS and the ETS logo are registered trademarks of Educational Testing Service (ETS). MEASURING THE POWER OF LEARNING is a trademark of ETS. 30141

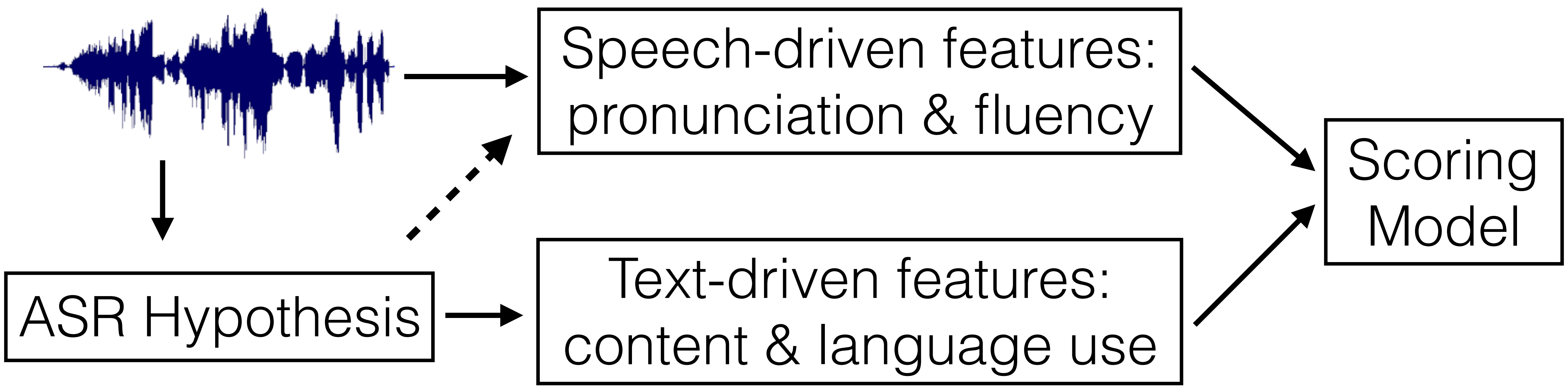
Automatically Scoring Spoken Responses

- Goal: assess test-taker's language proficiency

- *"Did the test-taker produce a [coherent, intelligible]() response that [addresses the question]()?"*

1

2



Copyright © 2017 by Educational Testing Service. All rights reserved. ETS and the ETS logo are registered trademarks of Educational Testing Service (ETS). MEASURING THE POWER OF LEARNING is a trademark of ETS. 30141

Corpus

- Large-scale English proficiency assessment; each test-taker answers 6 questions: 2 “general” and 4 “source-based”.
- 153,461 responses from 33,053 test-takers answering 147 questions.
- **General** (N=48). *Choose a recent event in your country that people want to talk about? Why are people interested in this event? Explain with specific details and reasons?*
- **Source-based** (N=99). *Listen to a fragment from a lecture in Psychology. Using points and examples from the lecture, explain the concept of groupthink.*
- Each response is 45-60 seconds (~100 words) and scored by professional raters on a 1–4 scale.

Copyright © 2017 by Educational Testing Service. All rights reserved. ETS and the ETS logo are registered trademarks of Educational Testing Service (ETS). MEASURING THE POWER OF LEARNING is a trademark of ETS. 30141

Automated Speech Recognition

- Uses the Kaldi Speech Recognition Toolkit.
- Acoustic model: 5 layer DNN, 13 MFCC-based features; Trigram language model.
- Trained on a proprietary corpus of 800 hours of similar speech from 8700 speakers with > 100 L1s.
- No speaker or question overlap.
- WER: ~30% on a similar corpus of spontaneous non-native speech. H-H inter-transcriber agreement: 15-20%.

Copyright © 2017 by Educational Testing Service. All rights reserved. ETS and the ETS logo are registered trademarks of Educational Testing Service (ETS). MEASURING THE POWER OF LEARNING is a trademark of ETS. 30141

Research Questions

- Does the model combining text-driven and speech-driven features outperform models based on a single set of features?
- If so, what information extracted from the acoustic signal is most crucial?
- Do the results depend on the type of question?

Copyright © 2017 by Educational Testing Service. All rights reserved. ETS and the ETS logo are registered trademarks of Educational Testing Service (ETS). MEASURING THE POWER OF LEARNING is a trademark of ETS. 30141

Speech-driven Features (N=33)

Name	Description	Examples	N
Speech Rate	General Speech fluency	Words/min, words/min after excluding leading/trailing pauses	3
Quality	Deviation of pronunciation from that expected of a proficient speaker	Average confidence score, average acoustic model score	6
Pausing	Pausing patterns in the response	Mean pause duration, mean number of pauses, pause-to-speech ratio	9
Timing	Patterns of durational variation of different segments	Proportion of vocalic intervals, standard devision of duration of consonantal intervals	9
Prosody	Rhythmic/Prosodic Patterns	Standard deviations of intervals between stressed syllables	6

Copyright © 2017 by Educational Testing Service. All rights reserved. ETS and the ETS logo are registered trademarks of Educational Testing Service (ETS). MEASURING THE POWER OF LEARNING is a trademark of ETS. 30141

Text-driven Features

- Speech-driven features that are dense vectors of continuous values.
- Text-driven features are sparse binary vectors; shown to be effective for scoring content in written responses†.
  - Lowercased word *n*-grams (*n*=1,2)
  - Lowercased character *n*-grams (*n*=2–5)
  - Syntactic dependency triples
  - Bins based on response length

†**Automatically Scoring Tests of Proficiency in Music Instruction.** *Proc. BEA Workshop, NAACL 2016*. Nitin Madnani, Aoife Cahill, and Brian Riordan.

Copyright © 2017 by Educational Testing Service. All rights reserved. ETS and the ETS logo are registered trademarks of Educational Testing Service (ETS). MEASURING THE POWER OF LEARNING is a trademark of ETS. 30141

Scoring Models

- Support Vector Regressor (scikit-learn) with an RBF kernel.
- Train/test: 70/30; learner hyper-parameters tuned via cross-validation on training set with MSE objective.
- 13 models trained for each of the 147 questions
  - **1 text-only**; **6 speech-only**: all speech features (1), each individual speech feature group (5); **6 combined**: text + each individual speech feature group (5), text + all speech features
- Evaluation metric: $R^2 \in [0, 1]$

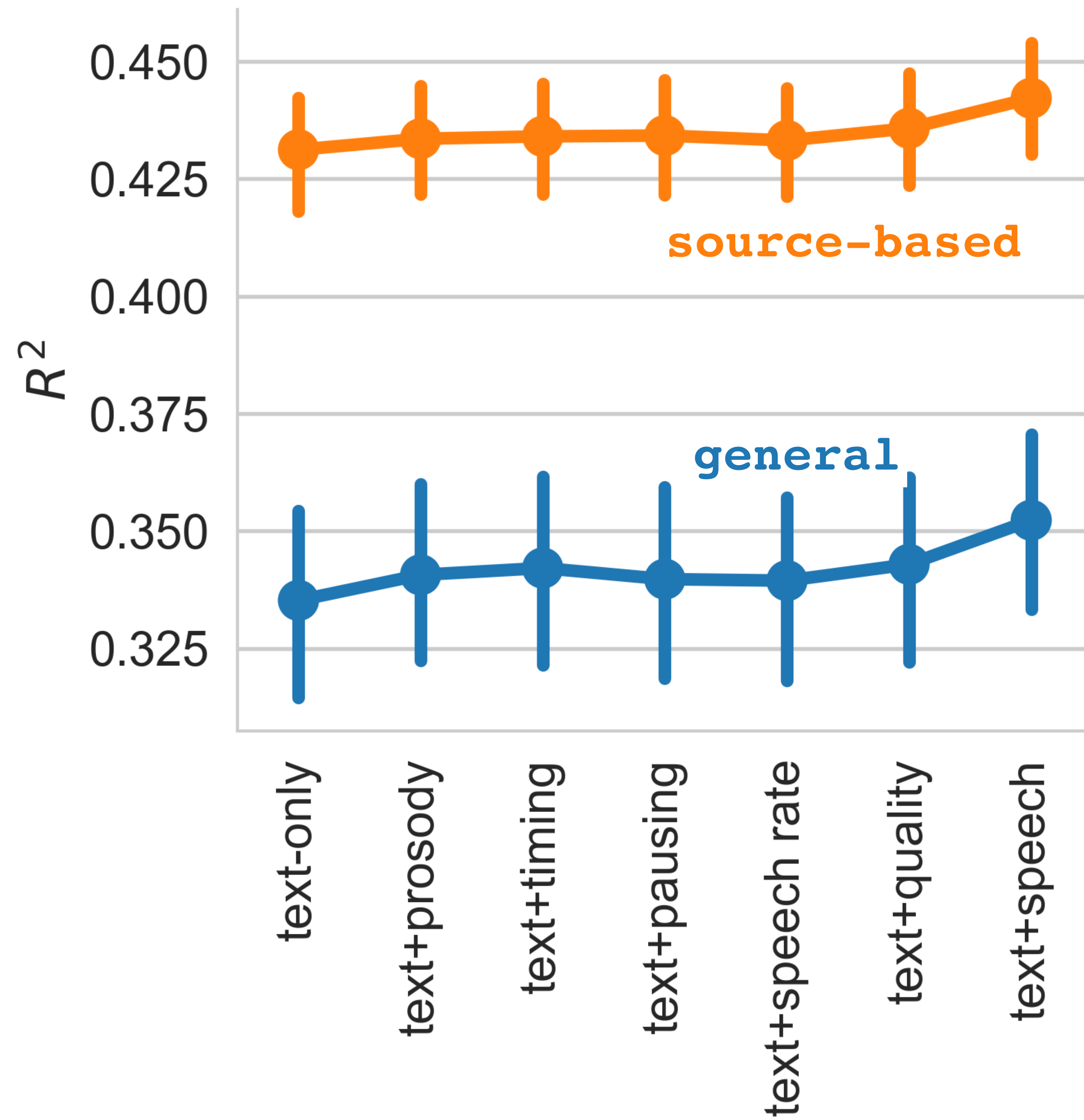
Copyright © 2017 by Educational Testing Service. All rights reserved. ETS and the ETS logo are registered trademarks of Educational Testing Service (ETS). MEASURING THE POWER OF LEARNING is a trademark of ETS. 30141

Model comparison: combined model

- All models performed better for source-based questions ($p < 0.0001$)
- Combined model outperforms speech-only and text-only models but only slightly (p=0.002)
- Text-only model outperforms speech-only model for source-based questions but not for general ones.

Model comparison: speech features

- text+any speech feature group does not outperform text-only.
- To obtain a small but significant improvement over text-only, we need to combine > 1 group of speech-driven features with text features.



Why such a small improvement?

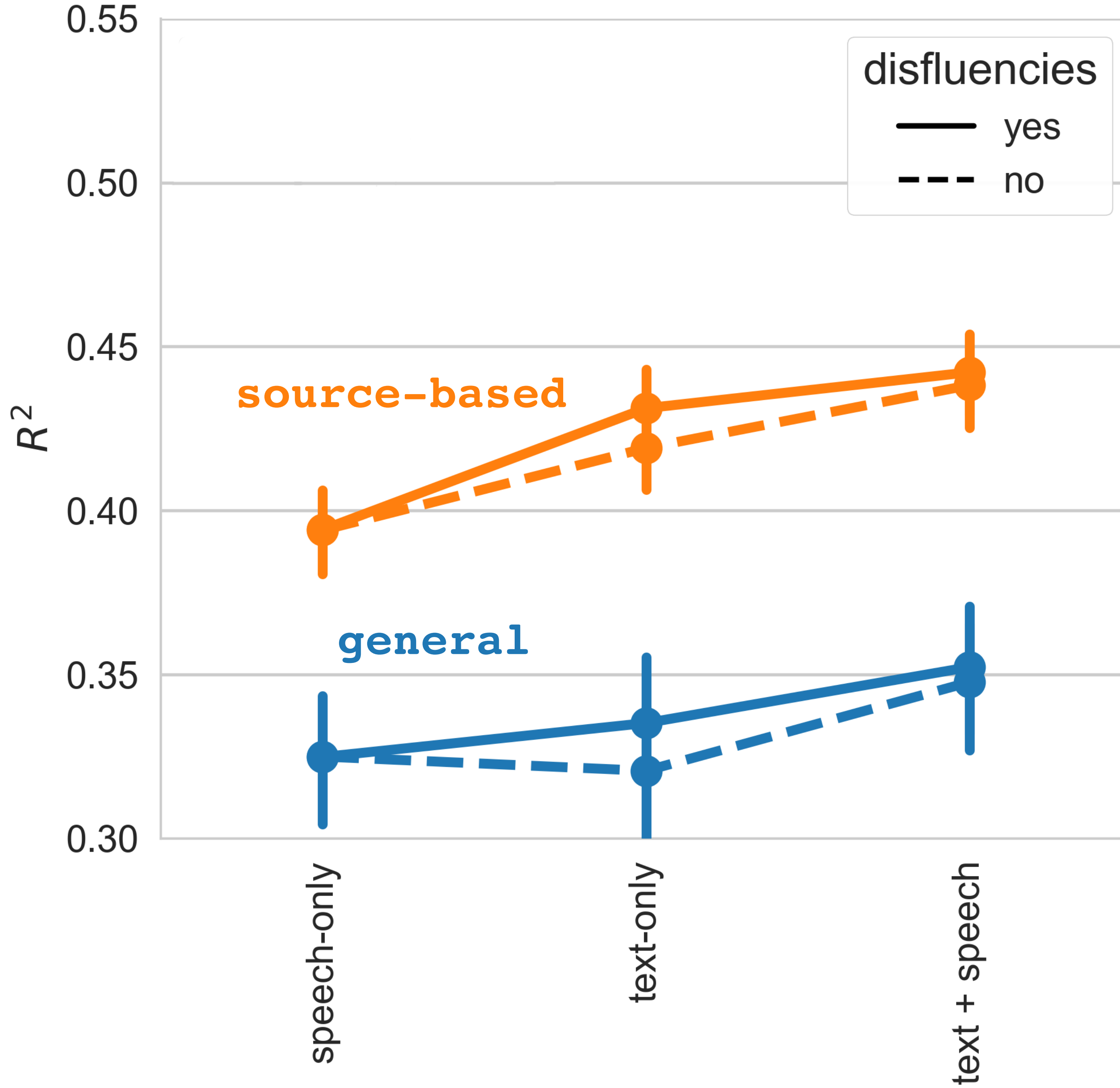
1. Speech features are ineffective. [Unlikely]
2. Poorly performing ASR. [Unlikely]
3. Ineffective combination. [Possible]
4. Significant overlap between information captured by the two modalities. [Likely]

Copyright © 2017 by Educational Testing Service. All rights reserved. ETS and the ETS logo are registered trademarks of Educational Testing Service (ETS). MEASURING THE POWER OF LEARNING is a trademark of ETS. 30141

Information Overlap

- Aspects of speaker proficiency captured by text and speech highly correlated, e.g., test-takers providing better content also pronounce better.

- Two sets of feature capture overlapping information, e.g., disfluencies & pausing patterns.



Conclusions - I

- Combination of speech & text features outperforms single modality with statistically significant but small improvement.

- Improvement in performance not due to any individual speech feature group.

- Text-driven features more effective for source-based questions than general ones. Surprisingly, similar results for speech-driven features.

Copyright © 2017 by Educational Testing Service. All rights reserved. ETS and the ETS logo are registered trademarks of Educational Testing Service (ETS). MEASURING THE POWER OF LEARNING is a trademark of ETS. 30141

Conclusions - II

- Text-only ASR hypothesis already captures a lot of information about speech.
- Adding further acoustic-based features may not always lead to substantial improvements, even when oral proficiency is critical.
- Our approach moderately effective but further research needed on how to obtain larger improvements, if any.

Copyright © 2017 by Educational Testing Service. All rights reserved. ETS and the ETS logo are registered trademarks of Educational Testing Service (ETS). MEASURING THE POWER OF LEARNING is a trademark of ETS. 30141

Questions?

nmadnani@ets.org

Copyright © 2015 by Educational Testing Service. All rights reserved. ETS and the ETS logo are registered trademarks of Educational Testing Service (ETS). MEASURING THE POWER OF LEARNING is a trademark of ETS. 30141

Effect of Question Type

- R^2 for best performing model between 0.06–0.51 for general and 0.20–0.56 for source-based.
- Sample size accounts for ~10–20% of variability (p < 0.001); significant but not main factor.
- Variation in ASR WER. Cannot measure directly but no significant effect for hypothesis length as a proxy.
- Additional analyses needed pertaining to question properties and test-taker characteristics.

Copyright © 2017 by Educational Testing Service. All rights reserved. ETS and the ETS logo are registered trademarks of Educational Testing Service (ETS). MEASURING THE POWER OF LEARNING is a trademark of ETS. 30141