An item response theory framework to evaluate automatic speech recognition systems against speech difficulty

Literature Review 2026.01.08

Jisoo Jang

Intro

- **Title**：An IRT framework to evaluate automatic speech recognition systems against speech difficulty
- **Journal(Year)** ：Computer Speech and Language 95 (2026)
- **Background**：results of ASR systems are usually assessed by aggregating ☐ varieties are ignored
- **Key Words**
    - Application of existing IRT theories to ASR
    - Plots (to show correlations between: sentence difficulties, system performance, speaker quality)

Table of Contents

- 1. About ASR Evaluation
- 2. Suggested Solution
- 3. IRT (Item Response Theory) in AI evaluation
- 4. IRT evaluation in ASR
- 5. RCC (Recognizer Characteristic Curves)
- 6. ASR Fingerprint
- 7. Discussion

1. About ASR Evaluation

Importance

- particular application ☐ which technique?
- to know advantages & limitations of existing techniques

Relying Aspects

- Dataset
- Accessing the quality of transcriptions

Limitation of existing method

– usually assessed by aggregating the results

– the variety of difficulties are ignored

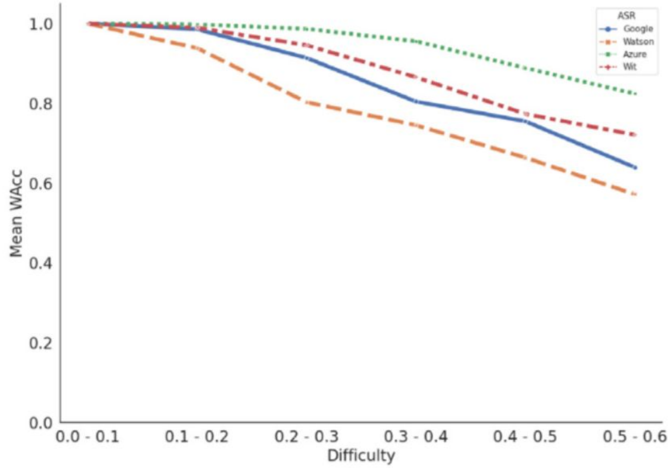
how can we overcome this limitation?

2. Suggested Solution

IRT (Item Response Theory)	RCC (Recognizer Characteristic Curve)	ASR Fingerprint
theory that evaluates – ability of ASR systems – difficulty of test speeches	**plot** x axis: speech difficulty y axis: ASR performance	**plot** x axis: sentence difficulty y axis: speaker quality color: ASR performance

$$E[r_{jki}|\theta_i,\delta_{jk}] = \frac{\gamma_{jki}}{\gamma_{jki}+\omega_{jki}} = \frac{1}{1+\left(\frac{\delta_{jk}}{1-\delta_{jk}}\right)\left(\frac{\theta_i}{1-\theta_i}\right)^{-1}}$$

$$E[\delta_{jk}|\varphi_k,w_j,a_j] = \frac{\alpha_{jk}}{\alpha_{jk}+\beta_{jk}} = \frac{1}{1+\left(\frac{\varphi_k}{1-\varphi_k}\right)^{a_j}\left(\frac{w_j}{1-w_j}\right)^{-a_j}},$$



3. IRT (Item Response Theory) in AI Evaluation

3.1 Definition

- A paradigm developed in Psychometrics
- usually proposed to model the probability of binary responses
- **Respondent's Ability** : Estimates the <u>ability of each AI system</u> (능력)
- **Item Difficulty** : Estimates the <u>difficulty of each test task</u>
    - Difficulty (난이도)
    - Discrimination (변별력)
    - Guessing (추측도)

– respondent's ability
– item's difficulties

3. IRT (Item Response Theory) in AI Evaluation

3.2 Formal Definition

“특정 테스트 항목 i에 대한 응답자 j의 응답은 확률 변수로 모델링되며, 이의 기댓값은 응답자 매개변수 벡터 Θj와 항목 매개변수 벡터 Δi 모두에 의존한다. 매개변수 값은 원칙적으로 알려지지 않지만, 테스트에서 관찰된 응답 모음으로부터 추정될 수 있다.”

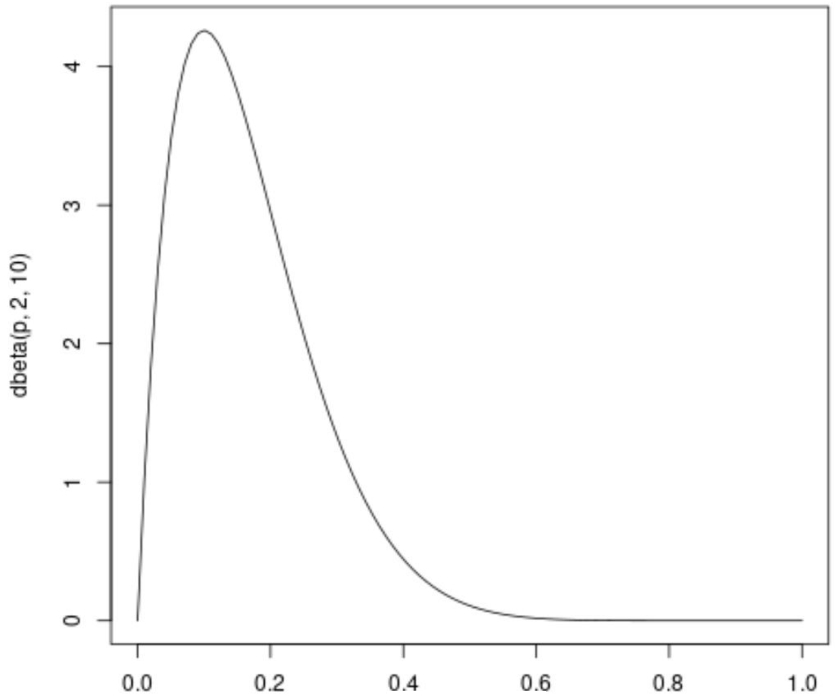
=

응시자(j)가 테스트(i)를 통과할지 말지는
1. 응시자의 능력 (θ_j)과
2. 항목의 난이도 (Δ_i)에
의존한다.
Θ, Δ은 테스트 결과로 추정 가능하다

3. IRT (Item Response Theory) in AI Evaluation

3.3 IRT Model β^3

- Chen et al., 2019
- The model used in this paper
- for bounded, continuous responses r_{ij} (음성 데이터)
- Reponses' probability follows beta distribution



– respondent's ability
– item's difficulties

?

r_{ij}

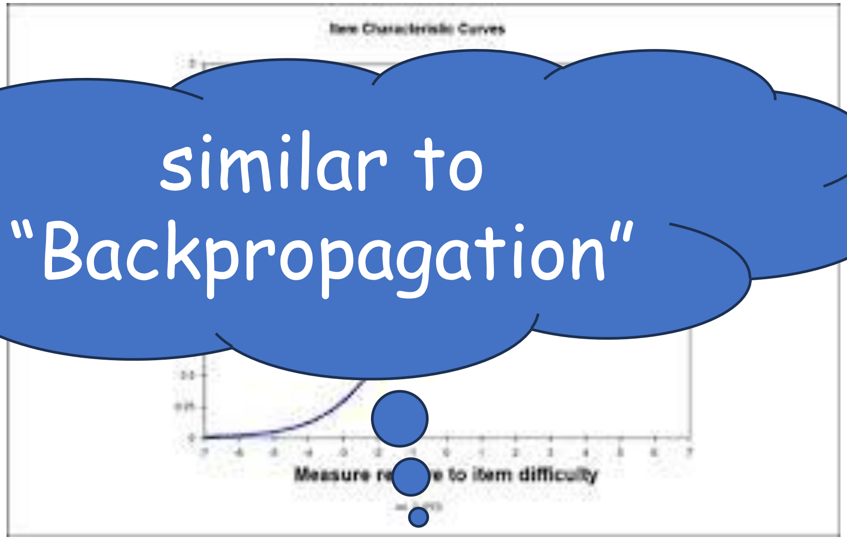
WER을 정규화한 값 등…

3. IRT (Item Response Theory) in AI Evaluation

3.4 ICC (Item Characteristic Curve)

- r_{ij} is calculated according to…

$$E[r_{ij}|\theta_j, \delta_i, a_i] = \frac{1}{1 + \left(\frac{\delta_i}{1-\delta_i}\right)^{a_i} \left(\frac{\theta_j}{1-\theta_j}\right)^{-a_i}}$$



similar to "Backpropagation"

- $r_{ij} \in [0, 1]$ is the response of respondent j to item i;
- θ_j is the ability of the respondent j;
- δ_i is the difficulty of the item i;
- a_i is the discrimination of the item i.

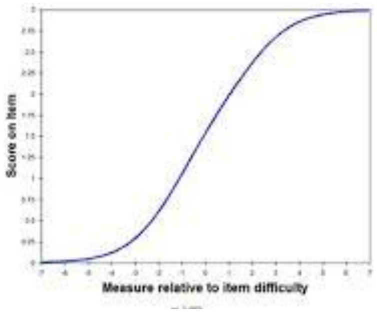
3. IRT (Item Response Theory) in AI Evaluation

3.4 ICC (Item Characteristic Curve)

- $parameter\ \theta, \delta$ inference:

초기
– ability parameter
– item's parameter

r_{ij}



via MLE, …

업데이트
– ability parameter
– item's parameter

마참내

– respondent's ability
– item's difficulties

3. IRT (Item Response Theory) in AI Evaluation

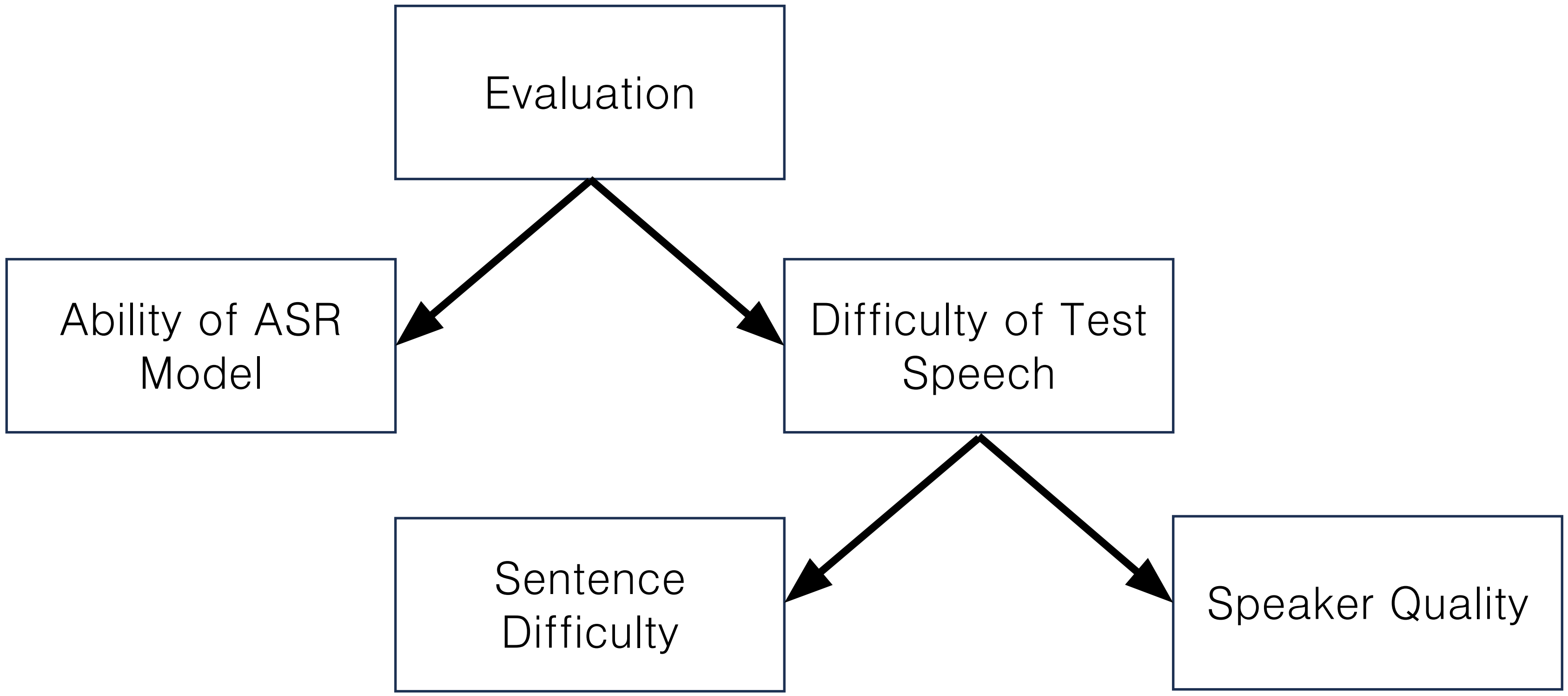
3.5 IRT in AI Domain

- Depending to AI Domain:
  - pool of respondents
  - pool of items
  - measuring method
- General Roadmap:
  - 1. choose pool of systems
  - 2. choose a benchmark of set of tasks
  - 3. evaluate the performance
  - 4. fit an appropriate IRT model (depends on domain of response)

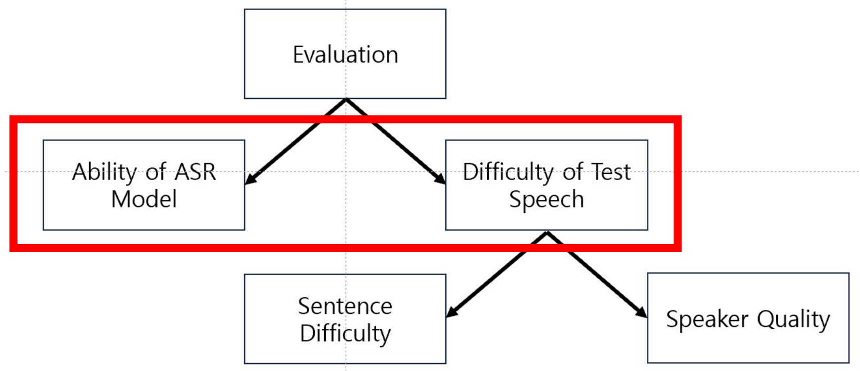
4. IRT evaluation in ASR

first level

second level



4. IRT evaluation in ASR



4.1 First Level IRT

- Objective: Ability of ASR Model, Difficulty of Test Speech

$$r_{jki} \sim Beta(\gamma_{jki}, \omega_{jki}),$$
$$\gamma_{jki} = \left(\frac{\theta_i}{\delta_{jk}}\right),$$
$$\omega_{jki} = \left(\frac{1-\theta_i}{1-\delta_{jk}}\right),$$
$$\theta_i \sim Beta(1,1), \delta_{jk} \sim Beta(1,1).$$

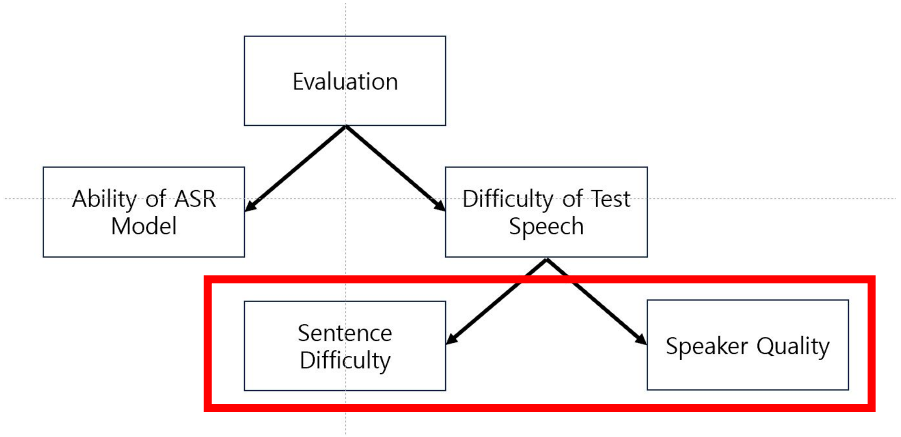
$$E[r_{jki}|\theta_i, \delta_{jk}] = \frac{\gamma_{jki}}{\gamma_{jki}+\omega_{jki}} = \frac{1}{1+\left(\frac{\delta_{jk}}{1-\delta_{jk}}\right)\left(\frac{\theta_i}{1-\theta_i}\right)^{-1}}.$$

r_{jki} Response
γ_{jki} Beta Parameter
ω_{jki} Beta Parameter
θ_i Ability of System
δ_{jk} Difficulty of Speech

High (better) responses are expected □ high-ability systems, easier speeches.

Low (worse) responses are expected □ low-ability systems, difficult speeches.

4. IRT evaluation in ASR



4.2 Second Level IRT

- Objective: Sentence Difficulty, Speaker Quality

$$\delta_{jk} \sim B(\alpha_{jk}, \beta_{jk}),$$

$$\alpha_{jk} = \left(\frac{\varphi_k}{w_j}\right)^{a_j},$$

$$\beta_{jk} = \left(\frac{1-\varphi_k}{1-w_j}\right)^{a_j},$$

$$\varphi_k \sim B(1,1), w_j \sim B(1,1), a_j \sim \mathcal{N}(1, \sigma_0^2).$$

$$E[\delta_{jk} | \varphi_k, w_j, a_j] = \frac{\alpha_{jk}}{\alpha_{jk} + \beta_{jk}} = \frac{1}{1 + \left(\frac{\varphi_k}{1-\varphi_k}\right)^{a_j} \left(\frac{w_j}{1-w_j}\right)^{-a_j}},$$

δ_{jk} Difficulty of Speech
α_{jk} Beta Parameter
β_{jk} Beta Parameter
φ_k Speaker's Quality
ω_j Sentence Difficulty
a_j Sentence Discrimination

4. IRT evaluation in ASR

4.3 Benchmarks

- Speaker: audio speeches were produced by adopting 4 TTS tools.
- Noise level: three levels of white noise was injected
- Total 7,500 speeches: (25 speakers) x (100 sentences) * (3 noise levels)

4.4 IRT fitting

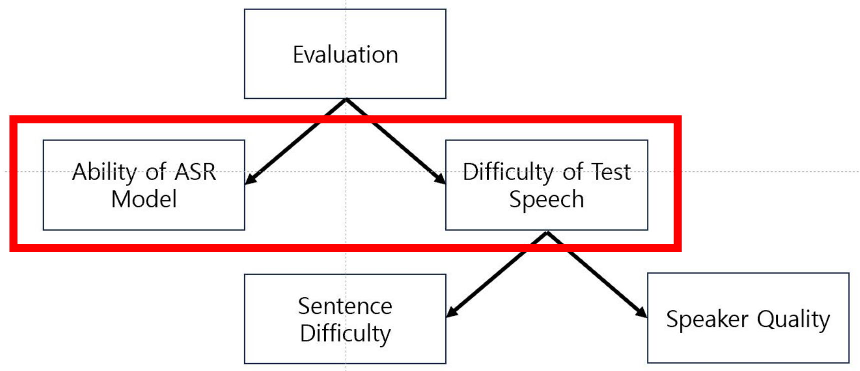
- Matrix for IRT model
- responses were defined as:

$$WAcc = 1 - \frac{S + D + I}{N}$$

in which:

- S is the number of substitutions;
- D is the number of deletions;
- I is the number of insertions;
- N is the number of words in the original sentence.

5. RCC

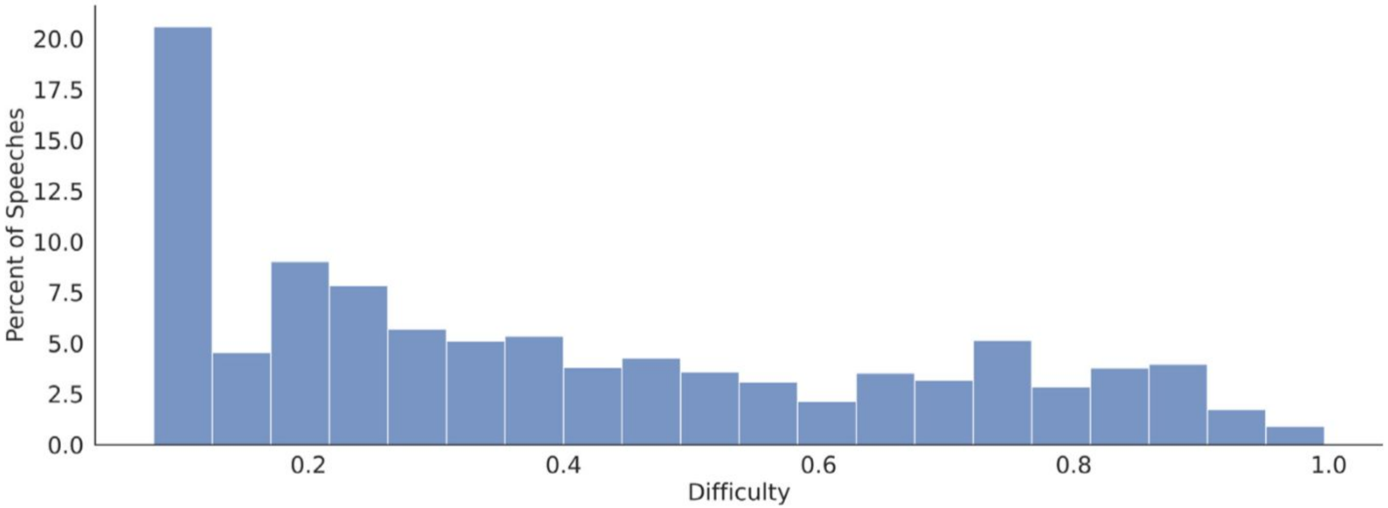


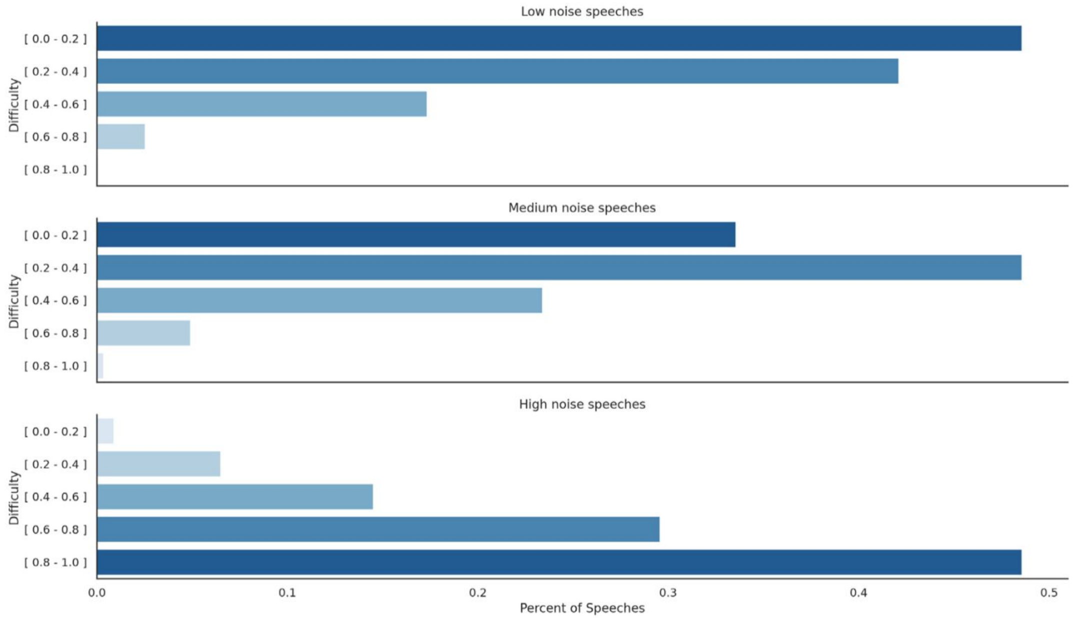
- Objective: Speech Difficulty □ ASR system's performance

$$R(\pi) = \int_{\delta} p(\delta) R(\pi|\delta) d\delta \qquad R(\pi, D) = \sum_{\delta} \hat{p}(\delta) \hat{R}(\pi, D|\delta)$$

$R(\pi)$	ASR systems's performance
$p(\delta)$	Difficulty Distribution
$R(\pi, D)$	System's performance conditioned by benchmark data D's difficulty
$\hat{p}(\delta)$	Frequency of Speeches in D
$\hat{R}(\pi, D\|\delta)$	system's performance averaged over the speeches with the same level of difficulty.

5. RCC





$R(\pi)$ ASR systems's performance

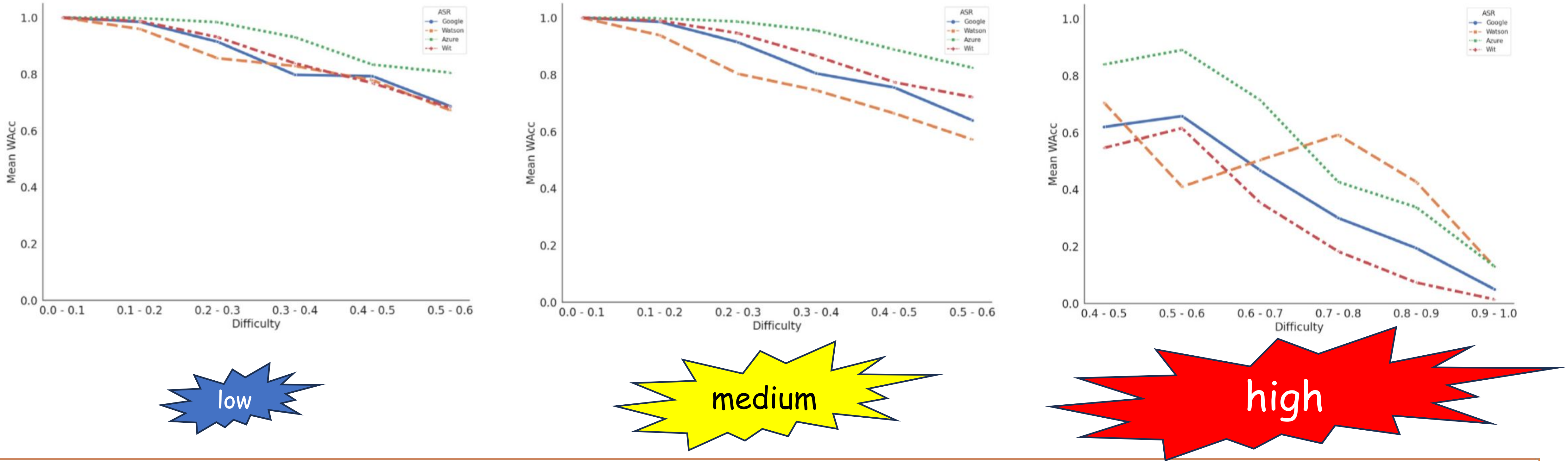
$p(\delta)$ Difficulty Distribution

$R(\pi, D)$ System's performance conditioned by benchmark data D's difficulty

$\hat{p}(\delta)$ Frequency of Speeches in D

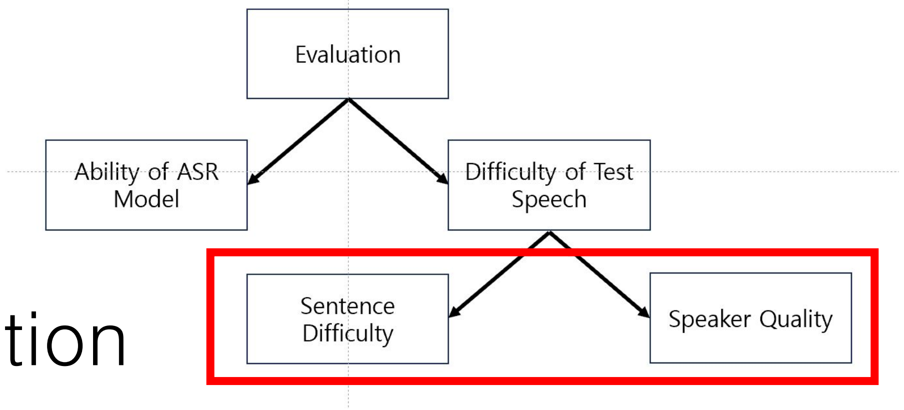
$\hat{R}(\pi, D|\delta)$ system's performance averaged over the speeches with the same level of difficulty.

5. RCC



$R(\pi)$	ASR systems's performance
$p(\delta)$	Difficulty Distribution
$R(\pi, D)$	System's performance conditioned by benchmark data D's difficulty
$\hat{p}(\delta)$	Frequency of Speeches in D
$\hat{R}(\pi, D\|\delta)$	system's performance averaged over the speeches with the same level of difficulty.

6. ASR Fingerprint



- Objective: Speech Difficulty Decomposition

$$R(\pi|\delta) = \int_{\omega,\phi} p(\omega,\phi|\delta) R(\pi|\omega,\phi) d(\omega,\phi)$$

Decompositio
n

$$R(\pi) = \int_{\delta} \int_{\omega,\phi} p(\delta) p(\omega,\phi|\delta) R(\pi|\omega,\phi) d(\omega,\phi) d\delta$$

Empirically estimated
as…

$$R(\pi, D) = \sum_{\delta} \sum_{\omega,\phi} \hat{p}(\delta) \hat{p}(\omega,\phi|\delta) \hat{R}(\pi, D|\omega,\phi)$$

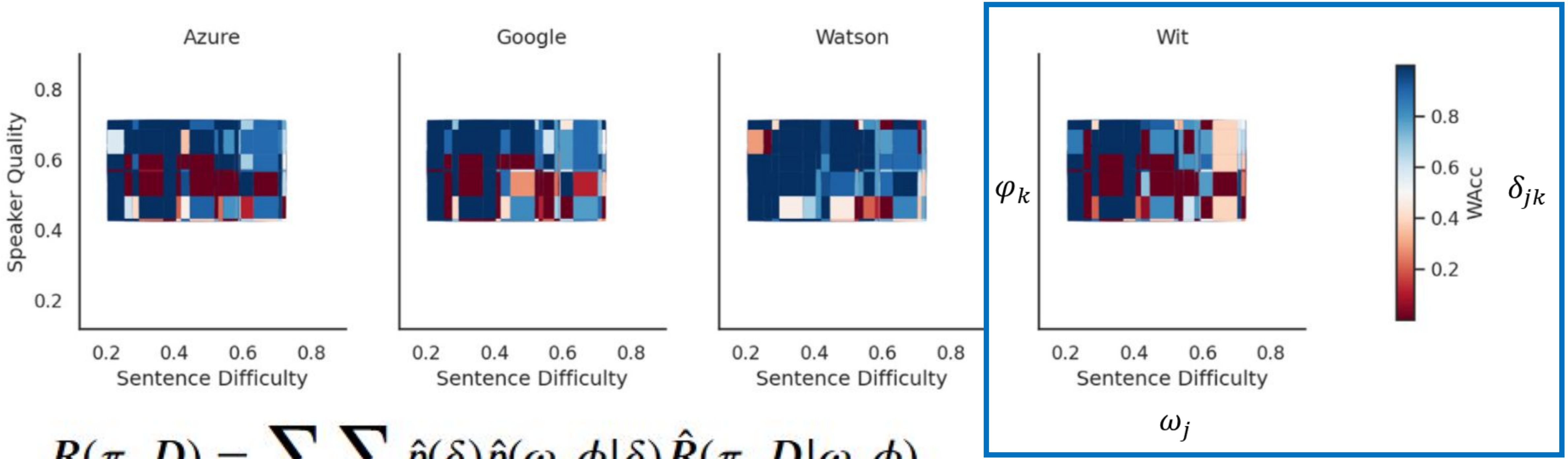
$R(\pi|\delta)$ Partial
Performance measure

δ_{jk} Difficulty of Speech
φ_k Speaker's Quality
ω_j Sentence Difficulty

6. ASR Fingerprint

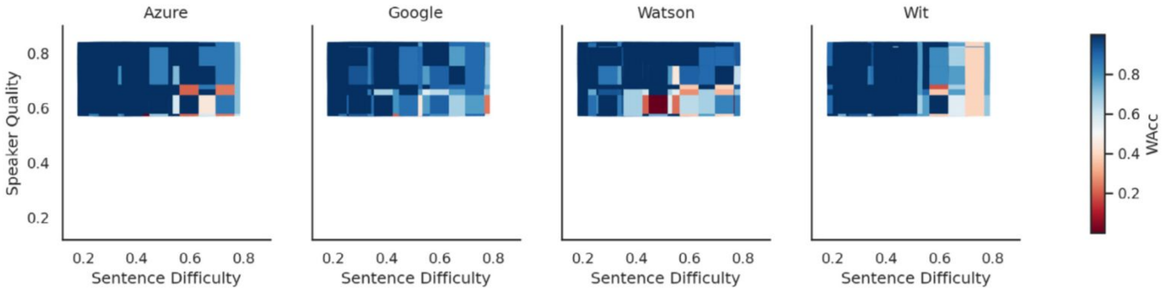
- for all noise level

δ_{jk} Difficulty of Speech
φ_k Speaker's Quality
ω_j Sentence Difficulty

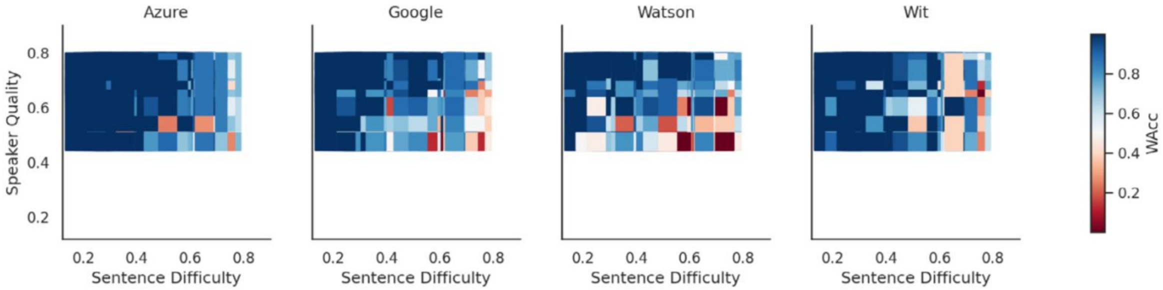


$$R(\pi, D) = \sum_{\delta} \sum_{\omega,\phi} \hat{p}(\delta)\hat{p}(\omega,\phi|\delta)\hat{R}(\pi, D|\omega,\phi)$$

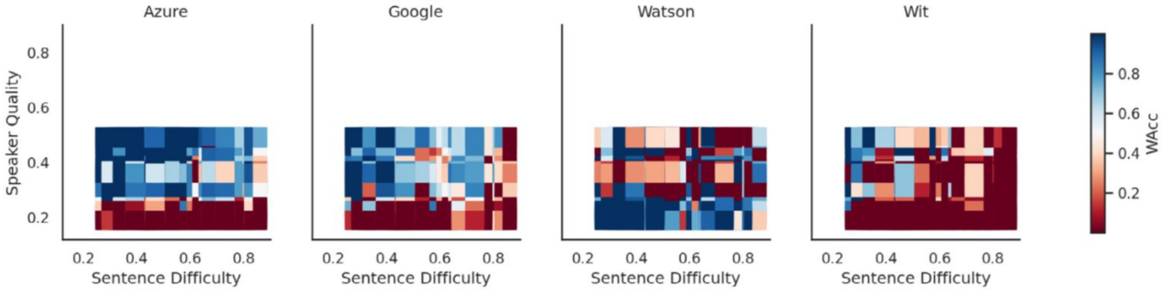
6. ASR Fingerprint



(a) Low noise level.



(b) Medium noise level.



(c) High noise level.

7. Discussion

- **Noise Injection**
  - Contribution: RCCs revealed impact of each noise injection on each ASR system's performances (medium is best)
  - Future work: richer noise types can be adopted
- **Human vs Synthetic test items**
  - Contribution: synthetic test items can produce diverse test items (scalability and cover many types)
  - Limitation: cannot express complex nuances of humans voices
  - Future work: Synthetic speech and human-record speech can be compared in terms of difficulty
- **Speech variability and data representativeness**
  - Contribution: many factors which makes speech diverse can be evaluated
  - Future work: larger and representative test benchmarks

Data Intelligence Lab

Thank you!

Q&A