Fairness in Speech-to-Text Algorithms

Anna Choi, April 22nd, 2024
A-Exam

Committee Members: Allison Koenecke (Chair), Matthew Wilkens, Marten Van Schijndel

Fairness in Speech-to-Text Algorithms

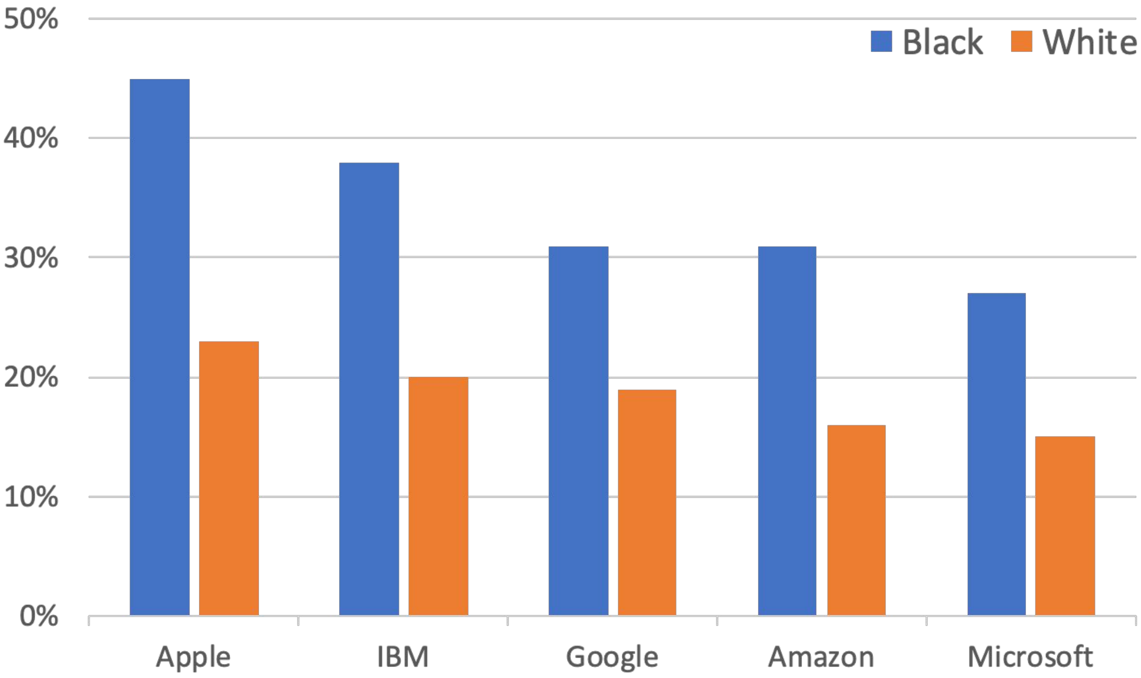
Comparing performance across different voice types

Fairness in **Speech-to-Text** Algorithms

Ability to translate spoken content into written words

Racial Disparity

Speech-to-text word error rates 2x worse for Black than white speakers



Koenecke, Allison, et al. "Racial disparities in automated speech recognition." Proceedings of the National Academy of Sciences 117.14 (2020): 7684-7689.

Evaluating Gender Bias in Speech Translation

Marta R. Costa-jussà[1], Christine Basta[1,2], Gerard I. Gállego[1]

Envisioning Equitable Speech Technologies for Black Older Adults

Robin N. Brewer
University of Michigan
Ann Arbor, MI, USA

Christina N. Harrington
Carnegie Mellon University
Pittsburgh, PA, USA

Courtney Heldreth
Google
Seattle, WA, USA

Language variation and algorithmic bias: understanding algorithmic bias in British English automatic speech recognition

Nina Markl

Word Error Rate (WER)

$$\text{WER} = \frac{\text{Substitution} + \text{Deletion} + \text{Insertion}}{\text{Number of Ground Truth Words}}$$

Word Error Rate (WER)

$$\text{WER} = \frac{\text{Substitution} + \text{Deletion} + \text{Insertion}}{\text{Number of Ground Truth Words}}$$

Ground truth: How are you today John
Transcription: How you a today Jones

Word Error Rate (WER)

$$\text{WER} = \frac{\text{Substitution} + \text{Deletion} + \text{Insertion}}{\text{Number of Ground Truth Words}}$$

Ground truth: How ~~are~~ you today John
Transcription: How you ~~a~~ today ~~Jones~~

Word Error Rate (WER)

$$\text{WER} = \frac{\text{Substitution} + \text{Deletion} + \text{Insertion}}{\text{Number of Ground Truth Words}}$$

Ground truth: How ~~are~~ you today John

Transcription: How you ~~a~~ today ~~Jones~~

WER = ⅗ = **0.6 (60 %)**

Fairness in Speech-to-Text Algorithms

Overview

1. Uncovering disparity
   a. d/Dhh project
   b. Aphasia project
2. Understanding components
   a. Speech data
   b. Text output
3. Future work

Quantification of Automatic Speech Recognition System Performance on d/Deaf and Hard of Hearing Speech

With Robin Zhao, Allison Koenecke, Anaïs Rameau
To be presented at COSM ALA 2024
To appear at The Laryngoscope

d/Deaf and Hard of Hearing Speech

Deafness is a severe hearing loss with very little to no functioning hearing.
Hard of hearing is a hearing loss that may have enough residual hearing to enable the use of an auditory device for assistance.

d/Deaf and Hard of Hearing Speech

Deafness is a severe hearing loss with very little to no functioning hearing.
Hard of hearing is a hearing loss that may have enough residual hearing to enable the use of an auditory device for assistance.

Characterized as:

Extremely slow, breathy or strained, monotone
Prolonged vowel production with results in distortion of syllables
Omission of final consonants

Variability by Speech Intelligibility, Onset of Hearing Loss, Communication Mode

Audio Data & Audit Target APIs

Speech Perception Assessment Laboratory (Univ. of Memphis)

Read speech of short story passages

24 d/DhH participants & 9 NH participants

484 audio files (291 d/Dhh & 153 NH)

OpenAI

Google

Microsoft Azure

aws

Mendel LL, Lee S, Pousson M, et al. Corpus of deaf speech for acoustic and speech production research. J Acoust Soc Am. 2017;142(1):EL102. doi:10.1121/1.4994288

Audit Results

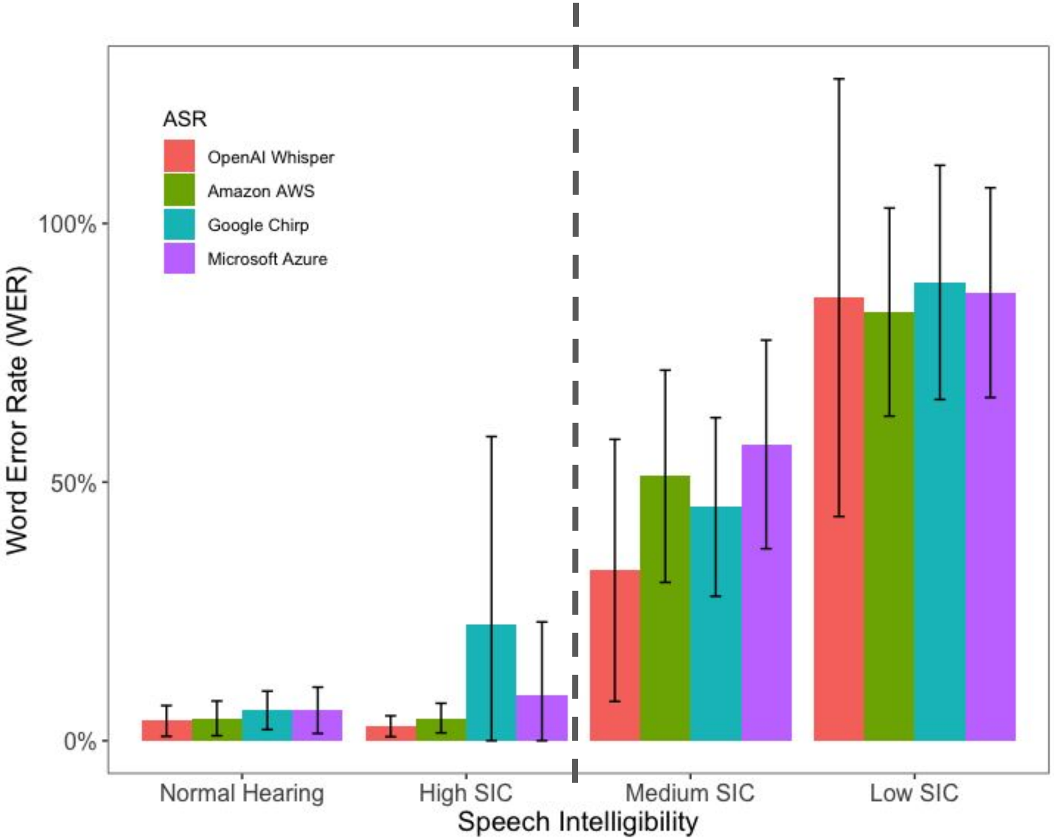
ASR Models	d/Deaf & Hard of Hearing	Normal Hearing
OpenAI Whisper	45.2 %	3.8 %
Google Chirp	55.7 %	5.9 %
Microsoft Azure	57.3 %	5.9 %
Amazon AWS	52.4 %	4.3 %
Average	52.7 %	5.0 %

Audit Results

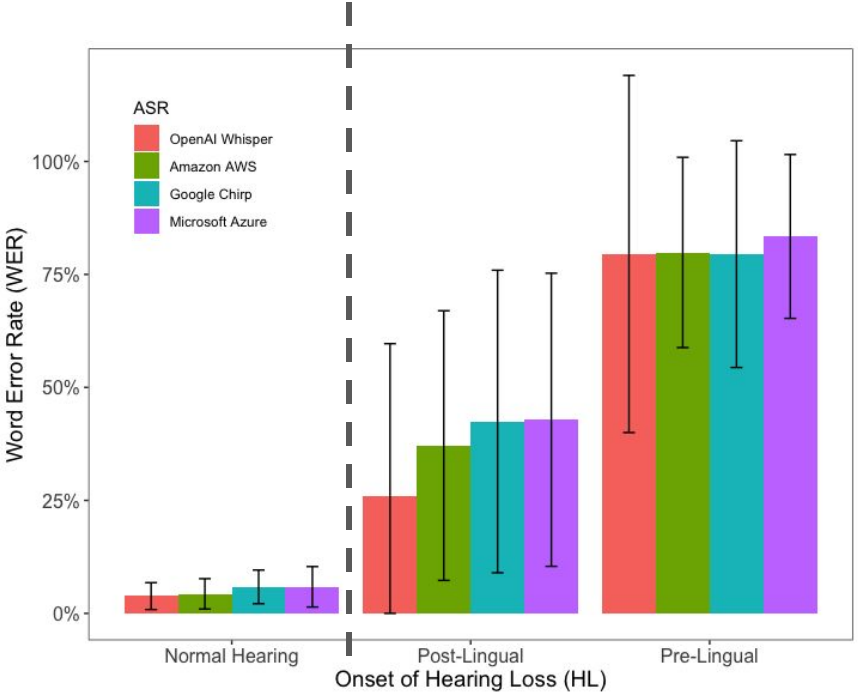
ASR Models	d/Deaf & Hard of Hearing	Normal Hearing
OpenAI Whisper	45.2 %	3.8 %
Google Chirp	55.7 %	5.9 %
Microsoft Azure	57.3 %	5.9 %
Amazon AWS	52.4 %	4.3 %
Average	52.7 %	5.0 %

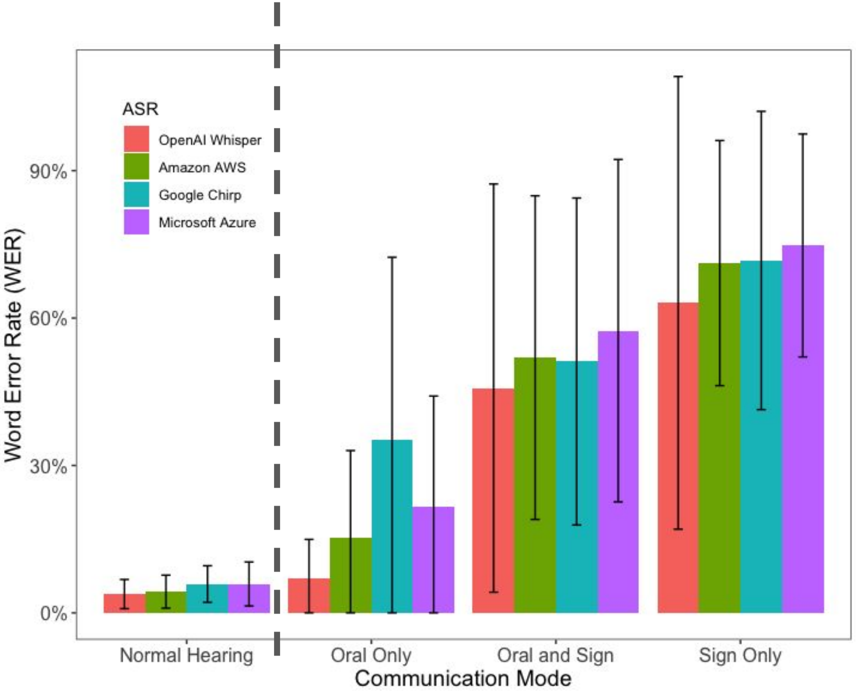
STT APIs perform 10X worse for d/Dhh

Audit Results



Audit Results

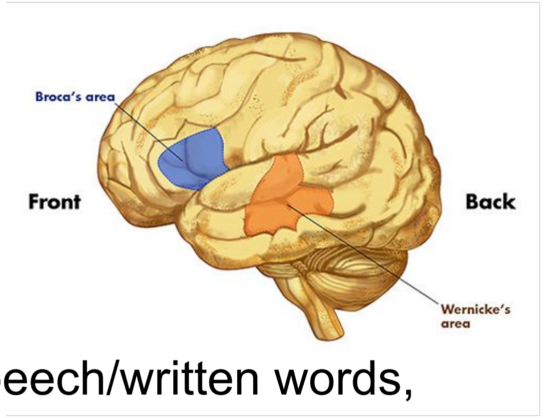




Quantification of Automatic Speech Recognition System Performance on Aphasia Speech

With Katelyn Mei (UW), Hilke Schellmann (NYU), Allison Koenecke, Mona Sloan (UVA)
In Preparation

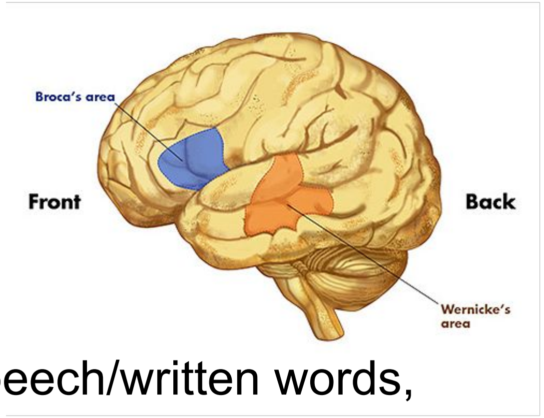
Aphasia Speech



Aphasia is a language disorder, caused by damage in a specific area of the brain that controls language

Difficulty with speaking/writing clearly, understanding speech/written words, remembering words

Aphasia Speech



Aphasia is a language disorder, caused by damage in a specific area of the brain that controls language

Difficulty with speaking/writing clearly, understanding speech/written words, remembering words

Non-fluent: difficulty initiating speech, no typical rhythm, short phrases with missing function words, long delays and pauses
Fluent: speaks smoothly with normal rhythm, nonsensical or made-up words, repetitions of sound patterns

Audio Data & Audit Target APIs

AphasiaBank (CMU)

551 Aphasia interviews & 347 non-Aphasia interviews

Average 38.8 seconds for Aphasia & 56.9 seconds for non-Aphasia

OpenAI

Google

Microsoft Azure

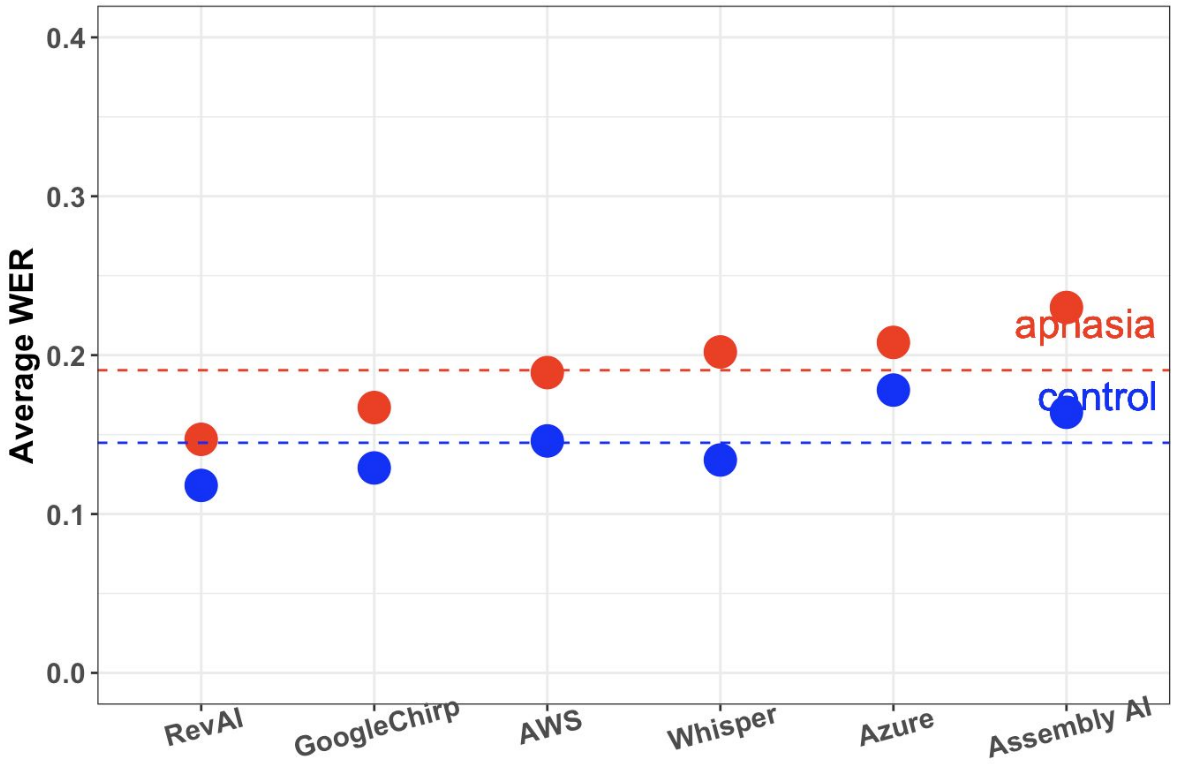
aws

rev

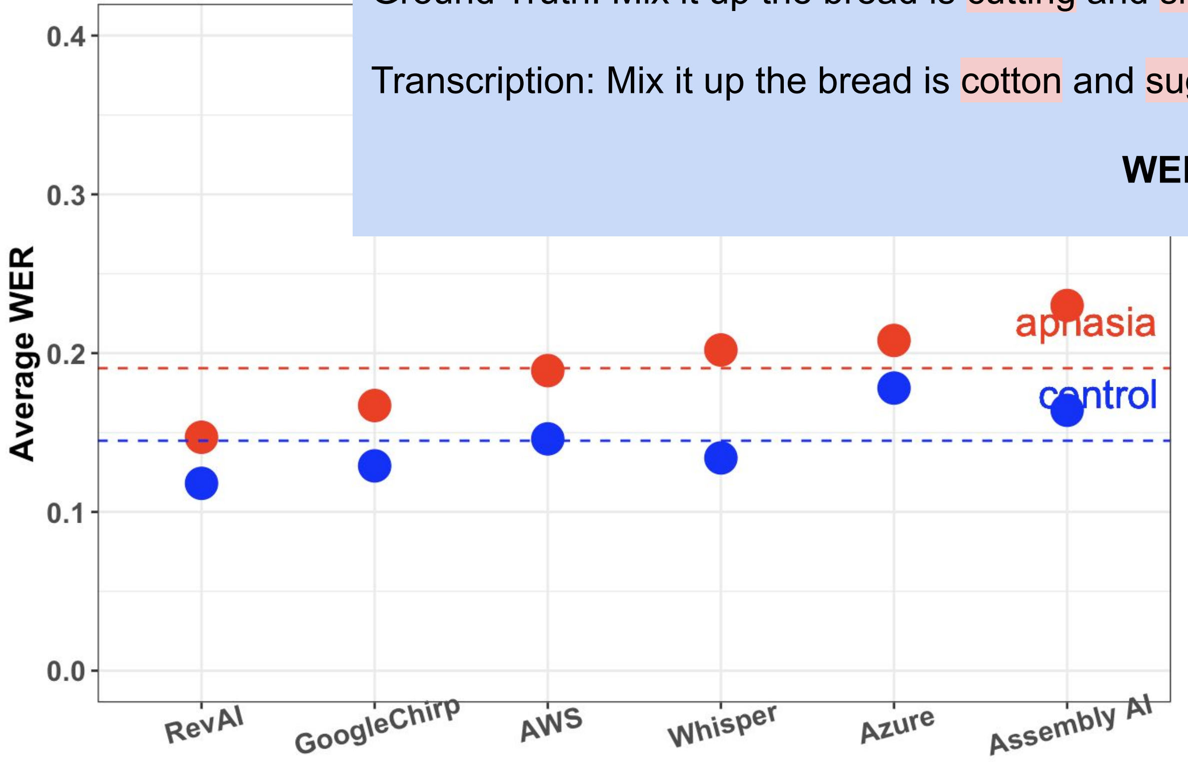
AssemblyAI

MacWhinney, B., Fromm, D., Forbes, M. & Holland, A. (2011). AphasiaBank: Methods for studying discourse. Aphasiology, 25,1286-1307

Audit Results



Audit Results

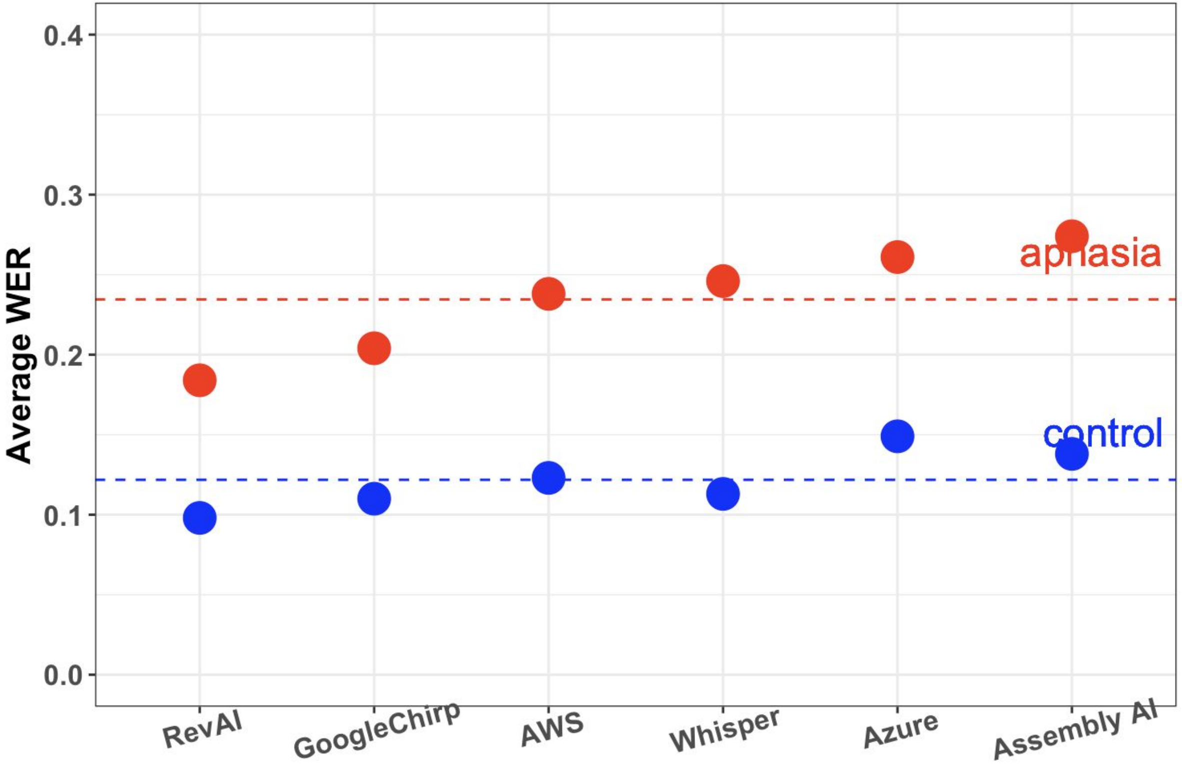


Ground Truth: Mix it up the bread is cutting and sharing

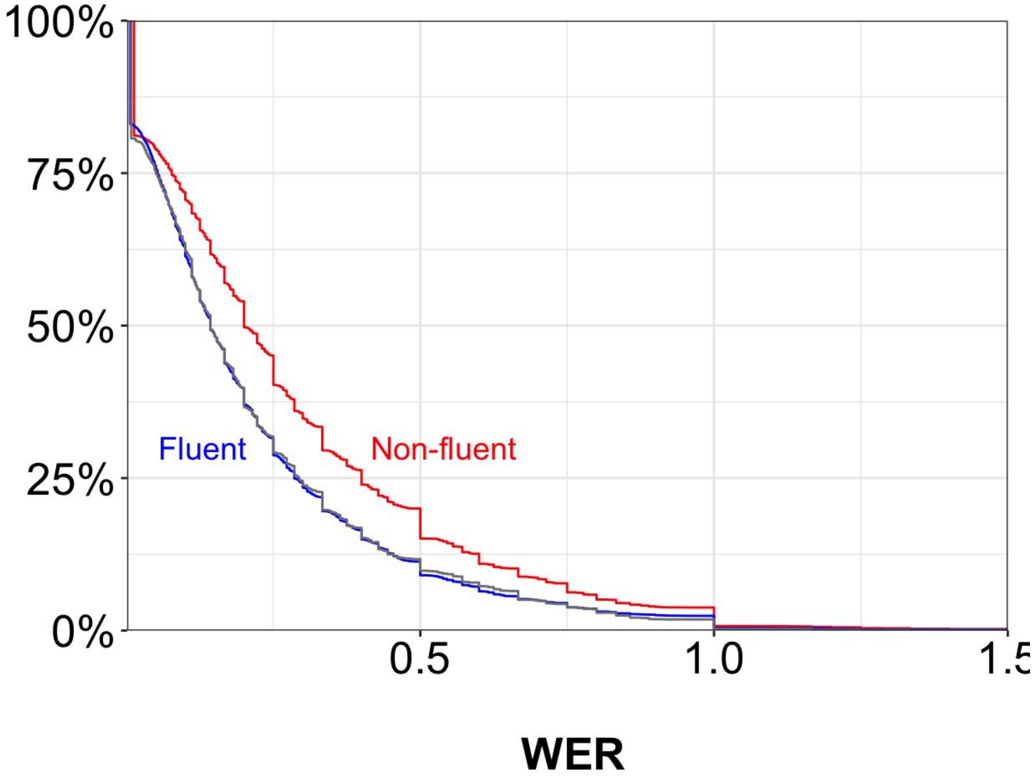
Transcription: Mix it up the bread is cotton and sugar

WER = 0.22

Audit Results - Demographically unmatched data



Audit Results

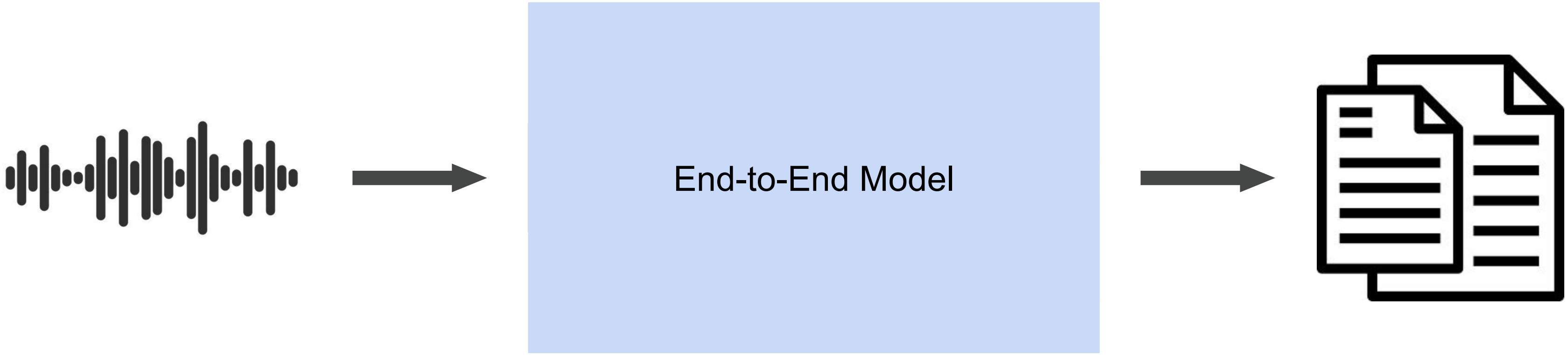


Fairness in Speech-to-Text Algorithms

Overview

1. Uncovering disparity
   a. d/Dhh project
   b. Aphasia project
2. Understanding components
   a. Speech data
   b. Text output
3. Future work

Speech-to-Text



Speech-to-Text

How should we collect diverse data to build a more inclusive STT?

Speech-to-Text

How should we evaluate model performance to ensure no further harms are caused?

Speech-to-Text

How should we build architectures that can mitigate bias?

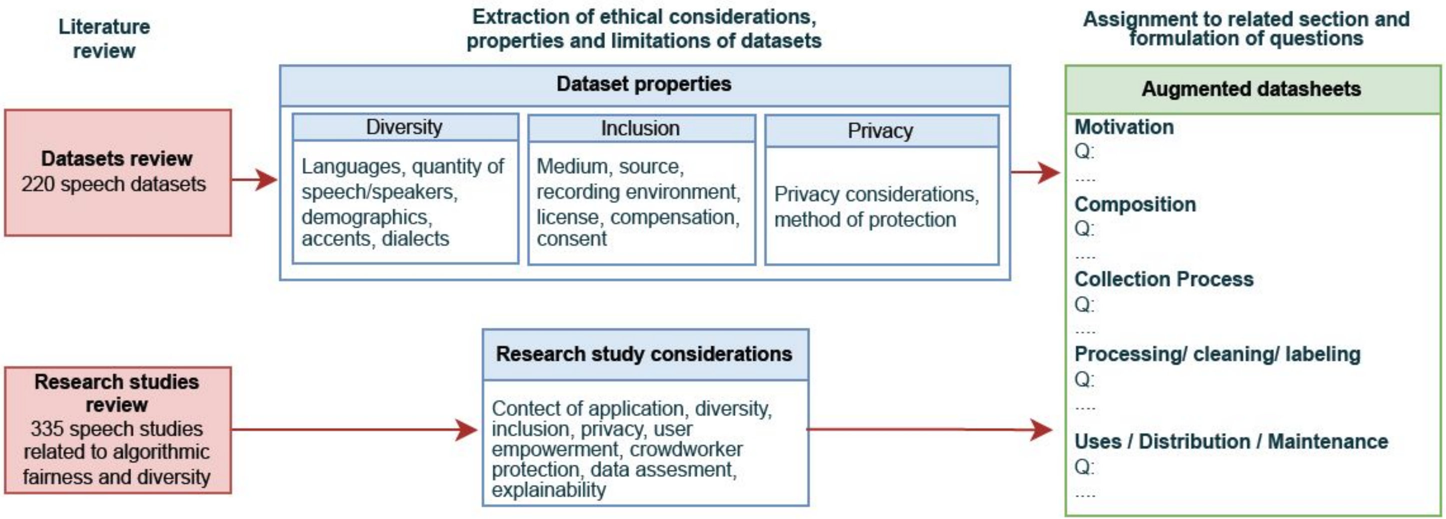
Speech-to-Text

Augmented Datasheets for Speech Datasets and Ethical Decision-Making

With Orestis Papakyriakopoulos, Jerone Andrews, Rebecca Bourke,
William Thong, Dora Zhao, Alice Xiang, Allison Koenecke
Presented at FAccT 2023
Presented at IC2S2 2023

Motivation

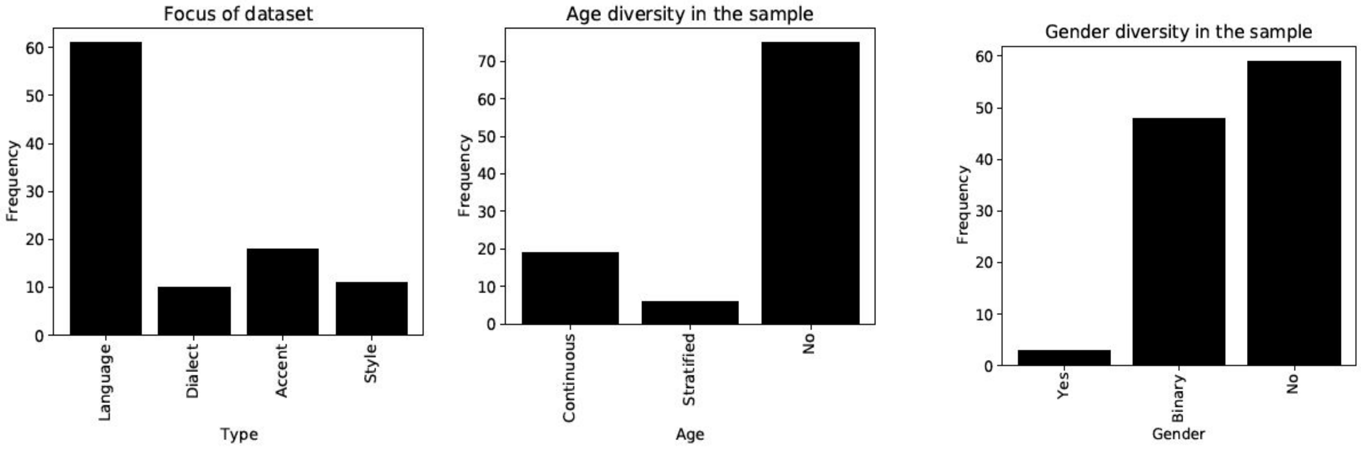
Building on Gebru et al.'s "Datasheets for Datasets" and augmenting for speech datasets specifically



Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for datasets. Commun. ACM 64, 12 (December 2021), 86–92. https://doi.org/10.1145/3458723

Motivation

Building on Gebru et al.'s "Datasheets for Datasets" and augmenting for speech datasets specifically



(a) Linguistic diversity **(b) Age diversity** **(c) Gender diversity**

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for datasets. Commun. ACM 64, 12 (December 2021), 86–92. https://doi.org/10.1145/3458723

Augmented Datasheets Sample Questions

1. Motivations

   Can the dataset be used to draw conclusions on read speech, spontaneous speech, or both?

2. Compositions

   Are there standardized definitions of linguistic subpopulations that are used to categorize the speech data? How are these linguistic subpopulations identified in the dataset and described in the metadata?

3. Collection Process

   Were all the data collected using the same technical methodology or setting, including the recording environment (e.g., lab, microphone) and recording information (e.g., sampling rate, number of channels)?

Augmented Datasheets Sample Questions

4. Preprocessing/Cleaning/Labeling

   Did the data collectors hire human annotators to transcribe the data? If so, how trained were the annotators in speech transcription for this context? How familiar were they with the corpus material, the vocabulary used, and the linguistic characteristics of different dialects and accents?

5. Uses/Distribution/Maintenance

   How are redactions performed on the dataset? Are personally identifiable information or sensitive information removed from only transcripts, audio censored from the speech data, or both?

Augmented Datasheets

A.1 Motivation

- What is the speech dataset name, and does the name accurately describe the contents of the dataset?
- Can the dataset be used to draw conclusions on read speech, spontaneous speech, or both?
- Describe the process used to determine which linguistic subpopulations are the focus of the dataset.

A.2 Composition

- How many hours of speech were collected in total (of each type, if appropriate), including speech that is not in the dataset? If there was a difference between collected and included, why? E.g., if the speech data are from an interview and the dataset contains only the interviewee's responses, how many hours of speech were collected in interviews from both interviewer and interviewee?
- How many hours of speech, number of speakers & words are in the dataset (by each type, if appropriate)?
- Are there standardized definitions of linguistic subpopulations that are used to categorize the speech data? How are these linguistic subpopulations identified in the dataset and described in the metadata?
- For any linguistic subpopulations identified in the dataset, please provide a description of their respective distributions within the dataset.
- How much of the speech data have corresponding transcriptions in the dataset?
- Does the dataset contain non-speech mediums (e.g. images or video)?
- Do speakers code switch or speak multiple languages, and if so, how is this identified in the data?
- Does the speech dataset focus on a specific topic or set of topics?
- Does the dataset include sensitive content that can induce different emotions (e.g., anger, sadness) that can cause the speakers to produce unusual pitch or tone deviating from plain speech?
- Does the dataset contain content that complies to the users' needs, or does it result in symbolic violence (the imposition of religious values, political values, cultural values, etc.)?

Augmented Datasheets

A.3 Collection Process

- What mechanisms or procedures were used to collect the speech data, e.g.: is the data a new recording of read speech or an interview? Or is it downloaded speech data from public speeches, lectures, YouTube videos or movies, etc.?
- Were all the data collected using the same technical methodology or setting, including the recording environment (e.g., lab, microphone) and recording information (e.g., sampling rate, number of channels)?
- Is there presence of background noise?
- For interviewer/interviewee speech data: during the interview process, did interviewers consistently ask questions that are "fair and neutral"?
- Have data subjects consented to the disclosure of the metadata in the dataset? Also, does the metadata include sensitive personal information such as disability status?

A.4 Preprocessing/cleaning/labeling

- When generating the dataset, was any background noise deleted or adjusted to make all recording qualities similar?
- Did the data collectors hire human annotators to transcribe the data? If so, how trained were the annotators in speech transcription for this context? How familiar were they with the corpus material, the vocabulary used, and the linguistic characteristics of different dialects and accents?
- If multiple transcription methods were used, how consistent were the annotators? How were transcripts validated?
- If the speech data include transcriptions, what software was used to generate the transcriptions (including, e.g., software used by human transcribers)? Are timestamps included in transcriptions? Are the alignments provided with the transcripts?
- Were transcription conventions (such as tagging scheme, treatment of hate speech or swear words, etc.) disclosed along with the corpus?
- Is additional coding performed, separate to transcriptions and tagging?

Augmented Datasheets

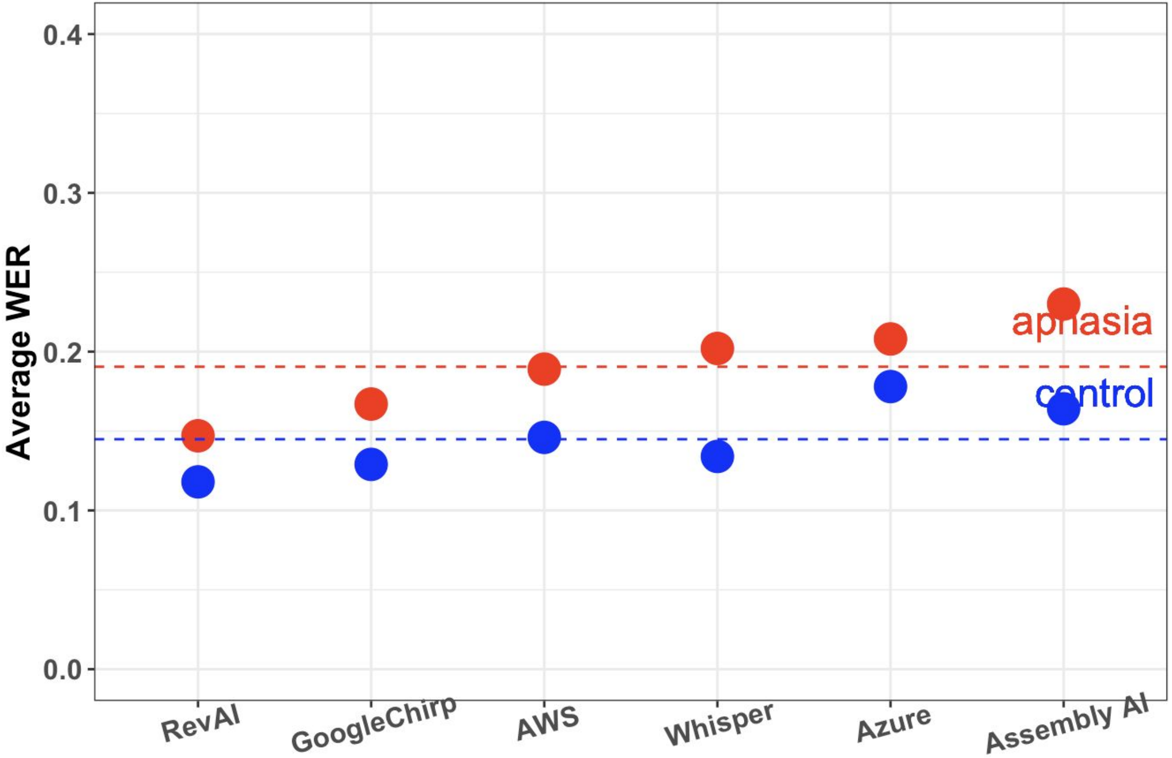
A.5 Uses / Distribution / Maintenance

- How are redactions performed on the dataset? Are personally identifiable information or sensitive information removed from only transcripts, audio censored from the speech data, or both?
- Is there any part of this dataset that is privately held but can be requested for research purposes?
- Is there a sample dataset distributed? If so, how well does the sample represent the actual dataset? Do they include all forms of speech included in the dataset? How big is the sample?
- Aside from this datasheet, is other documentation available about the data collection process (e.g., agreements signed with data subjects and research methodology)?

Speech-to-Text

How should we define ‘Ground Truth’?
Is WER an accurate measure of STT performance?

Aphasia Audit Results



Aphasia Data Standardization Method

Type of Transcript	Version	Main cleaning Steps involved in each version			
		Remove Fillers	Remove Fragments in Ground Truth	Remove Repeated Words	Remove Repeated Phrases
Ground truth	V1	☑	☐	☐	☐
Ground truth	V1+	☑	☑	☐	☐
Ground truth	V2	☑	☑	☑	☐
Ground truth	V3	☑	☑	☑	☑
ASR	V1	☑	☐	☐	☐
ASR	V1+	☑	☑	☐	☐
ASR	V2	☑	☑	☑	☐
ASR	V3	☑	☑	☑	☑

Original
H- h- he he wanted to, they were um, ball they were having a ball

V1
H- h- he he wanted to, they were, ball they were having a ball

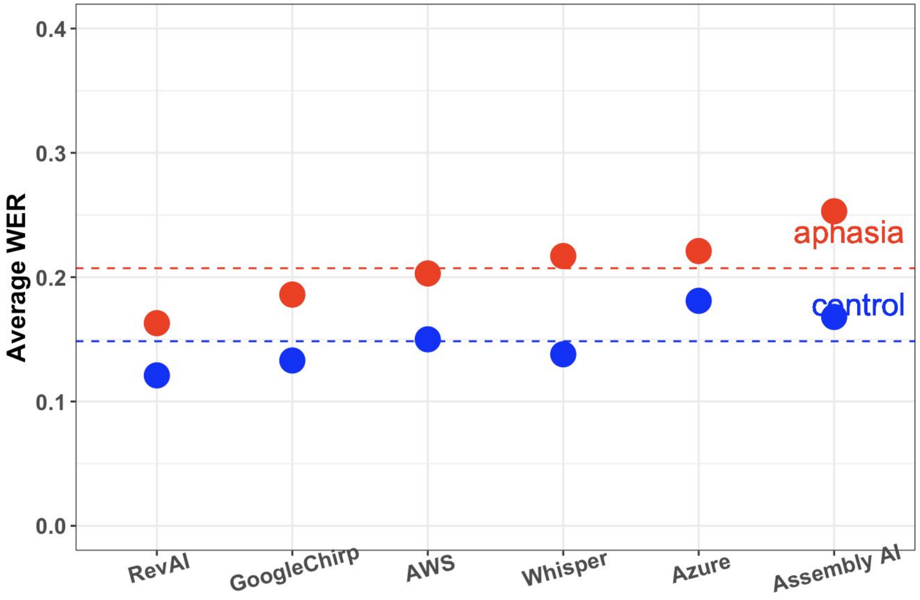
V1+
He he wanted to, they were, ball they were having a ball

V2
He wanted to, they were, ball they were having a ball

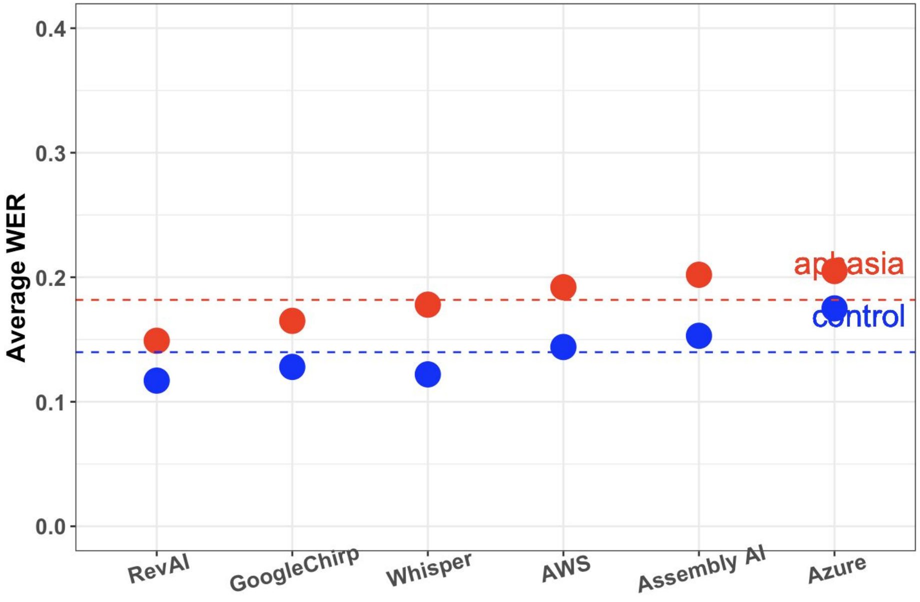
V3
He wanted to, they were having a ball

Aphasia Audit Results

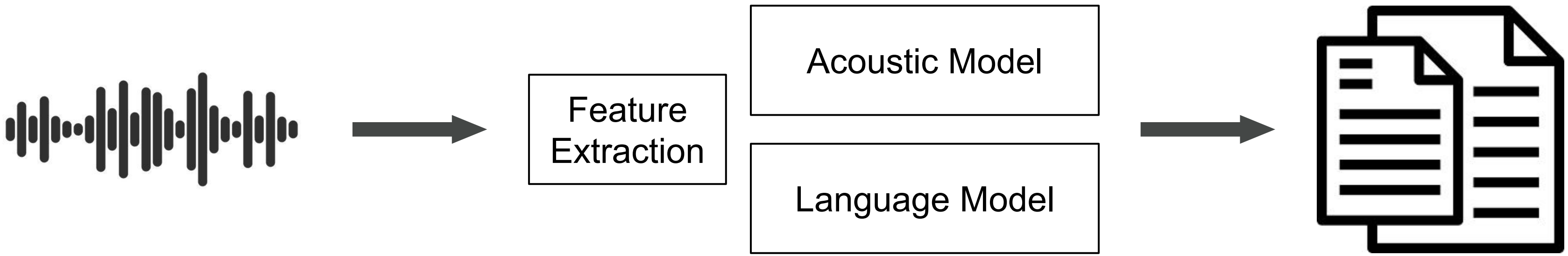
V1 standardization



V3 standardization



Speech-to-Text



Speech-to-Text

Careless Whisper: Speech-to-Text Hallucination Harms

With Allison Koenecke, Katelyn Mei (UW), Mona Sloan (UVA), Hilke Schellmann (NYU)
To be presented at FAccT 2024

Aphasia Audit Findings

What is Hallucination?

Undesirable generated text that is not present in the given input

Aphasia Audit Findings

What is Hallucination?

Undesirable generated text that is not present in the given input

Ground Truth	OpenAI Whisper
Someone had to run and call the fire department to rescue both the father and the cat.	Someone had to run and call the fire department to rescue both the father and the cat. All he had was a smelly old ol’ head on top of a socked, blood-soaked stroller.

Aphasia Audit Findings

What is Hallucination?

Undesirable generated text that is not present in the given input

Ground Truth	OpenAI Whisper
Everybody in the truck, the whole family, just waving and yelling. My goodness.	Everybody in the truck, the whole family, just waving and yelling. My goodness. That was pretty, extremely barbaric.

Aphasia Audit Findings

What is Hallucination?

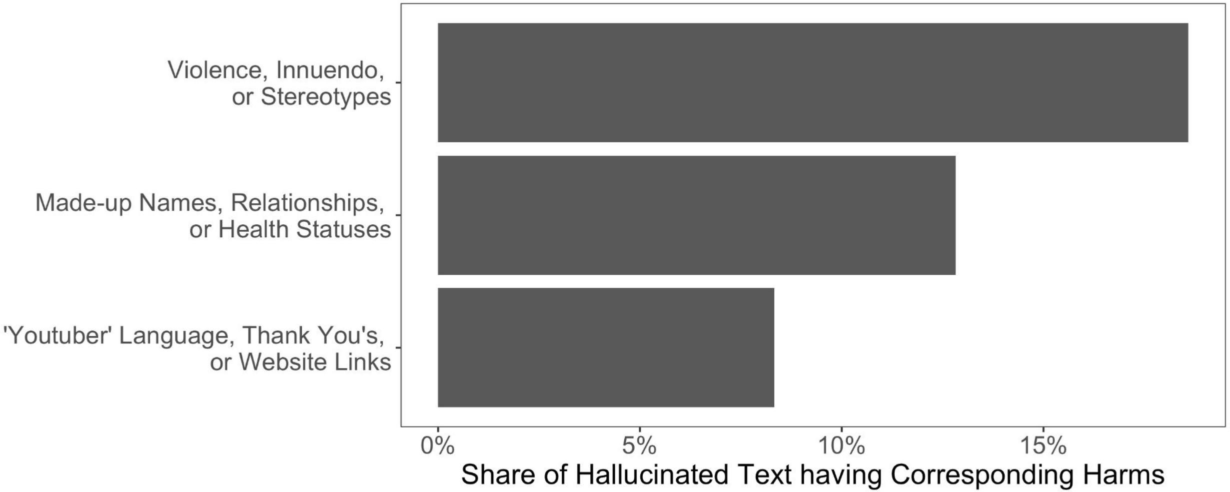
Undesirable generated text that is not present in the given input

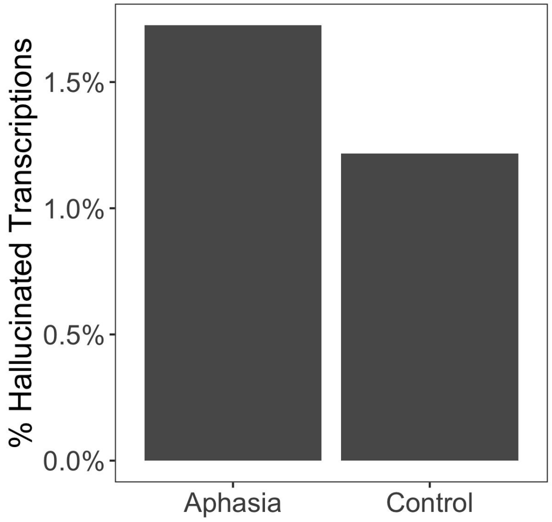
Ground Truth	OpenAI Whisper
Cinderella danced with the prince and…	Cinderella danced with the prince and… Thank you for watching!

Aphasia Audit Findings

What is Hallucination?

Undesirable generated text that is not present in the given input





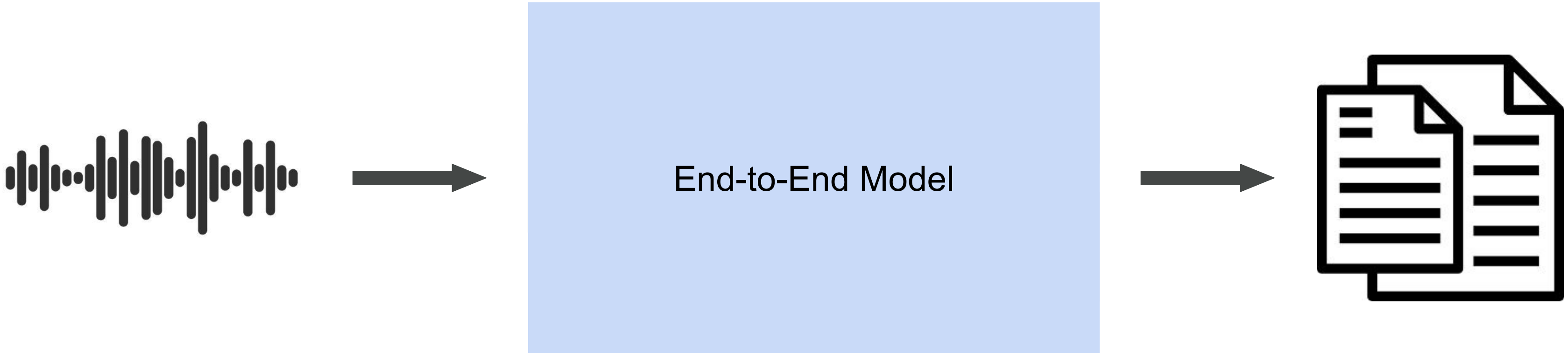
Fairness in Speech-to-Text Algorithms

Overview

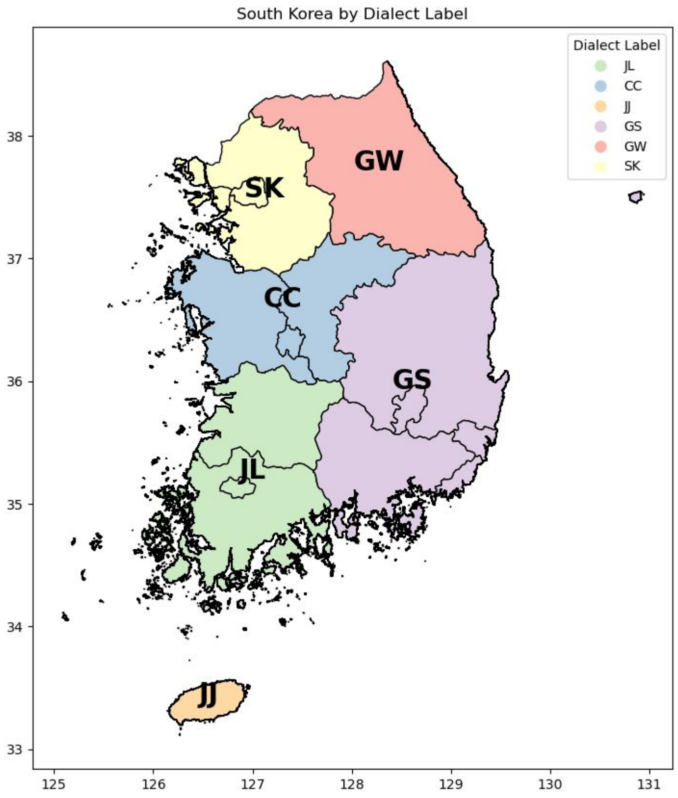
1. Uncovering disparity
   a. d/Dhh project
   b. Aphasia project
2. Understanding components
   a. Speech data
   b. Text output
3. Future work

Fairness in Korean Speech-to-Text Algorithms

Speech-to-Text

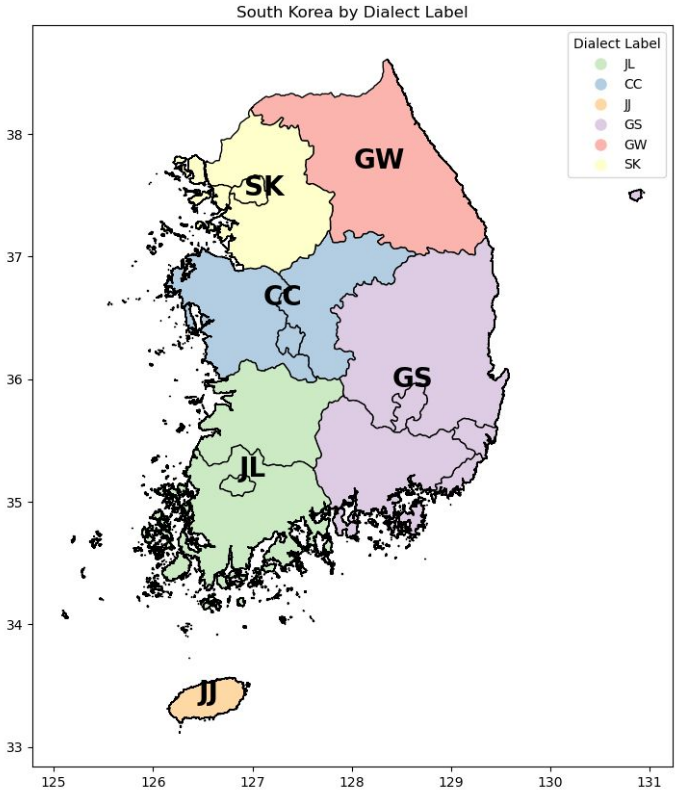


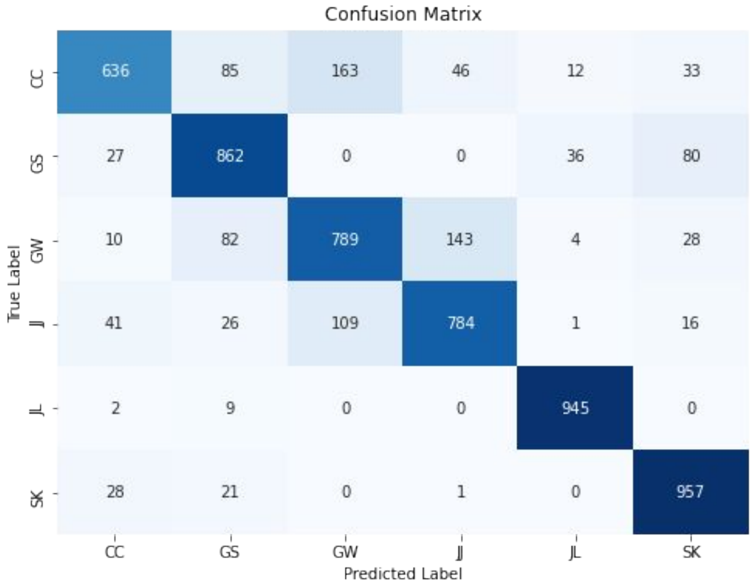
Korean Dialects



Regional dialects of Korean
Standard Korean
Gangwon
Chungcheong
Gyeongsang
Jeolla
Jeju

Korean Dialects





Speech-to-Text

Auditing Korean Speech Datasets for Dialectal Fairness in Speech-to-Text Applications, IC2S2 2023

Dataset Audit

AI Hub from Korean government

0.5 TB, 2,000 speakers, 3,000 hours of speech for each dialect

Qualitative Audit

1. Speech Collection
   a. Different speaker numbers
   b. Spontaneous vs Read speech
2. Transcription
   a. Formatting errors
   b. Grammar/Spelling errors

https://aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=data&dataSetSn=118

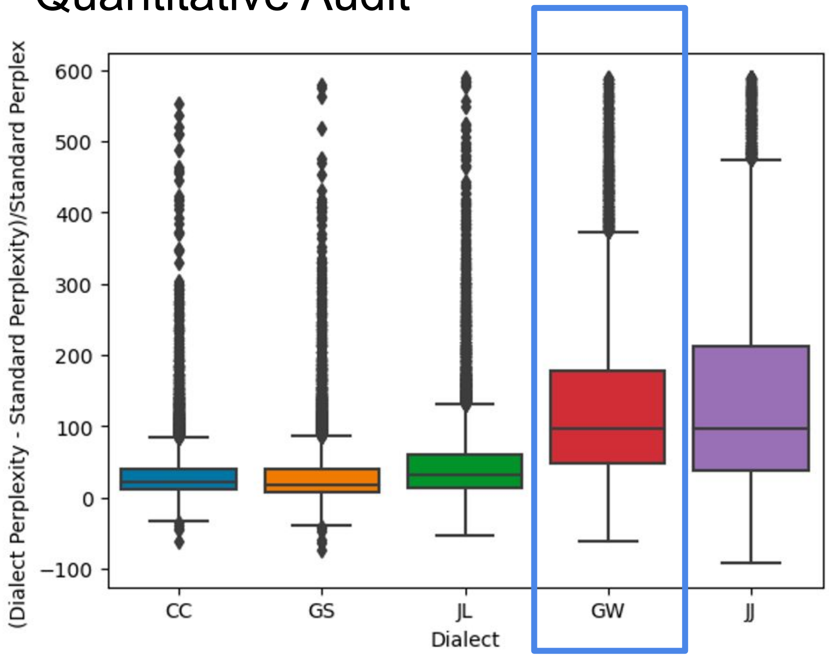
Dataset Audit

AI Hub from Korean government

5 TB, 2,000 speakers, 3,000 hours of speech for each dialect

Perplexity measure using KoGPT2

Quantitative Audit



https://aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=data&dataSetSn=118

Speech-to-Text

Would Korean STT Models perform just as well for Korean dialects as for the Standard Korean?

Audio Data & Audit Target APIs

AI Hub from Korean government

0.5 TB, 2,000 speakers, 3,000 hours of speech for each dialect

OpenAI

Google

Microsoft Azure

NAVER

kakao

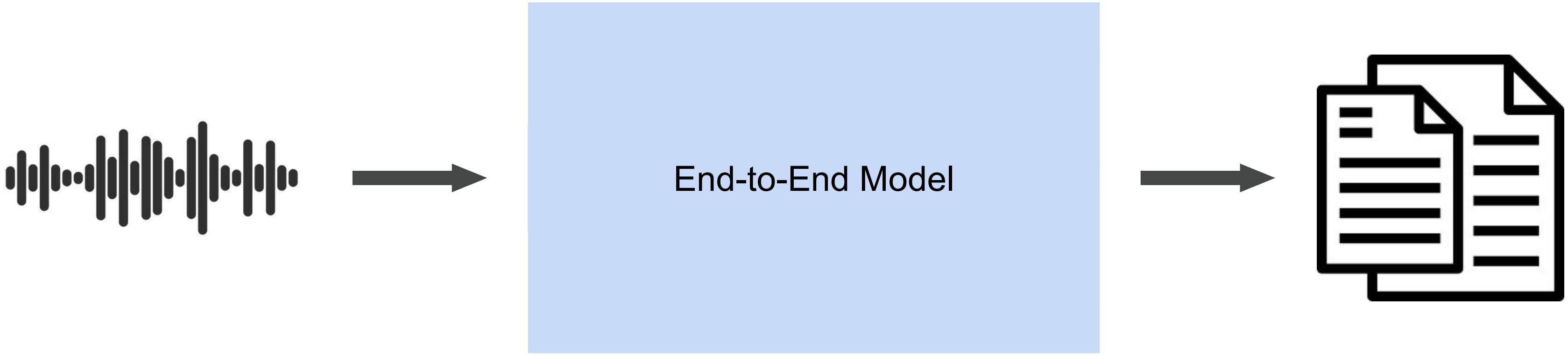
https://aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=data&dataSetSn=118

Speech-to-Text

What methods can I take on building a dialect-specific Korean STT Model?

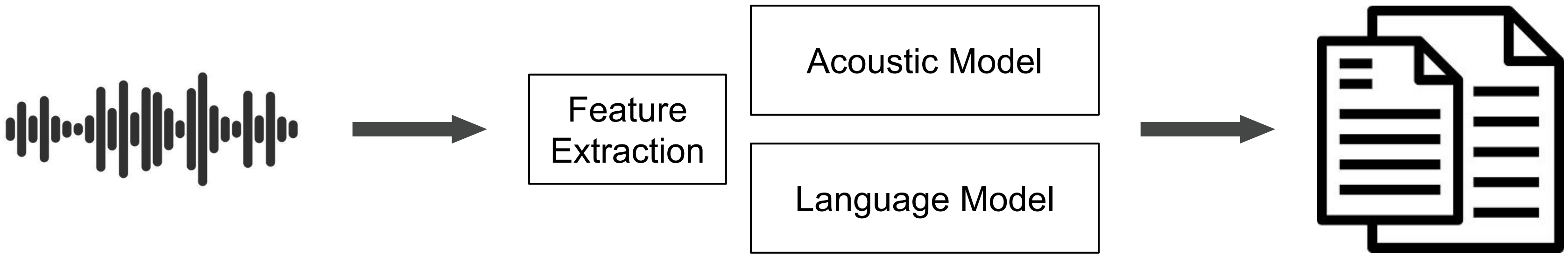
Fine-tuning for Korean Dialects

Fine-tuning is often used for low-resource languages or subgroups



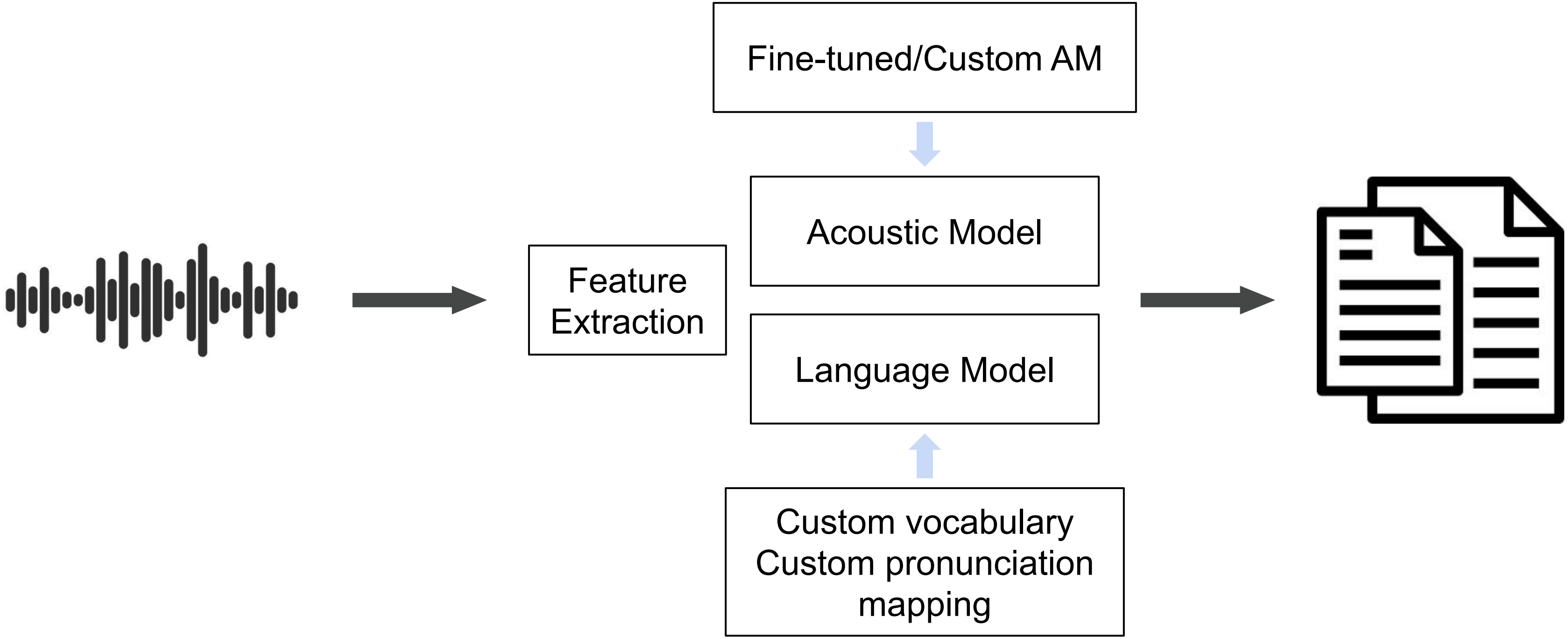
Fine-tuning for Korean Dialects

Fine-tuning is often used for low-resource languages or subgroups



Fine-tuning for Korean Dialects

Fine-tuning is often used for low-resource languages or subgroups



Fine-tuning for Korean Dialects

Fine-tuning is often used for low-resource languages or subgroups

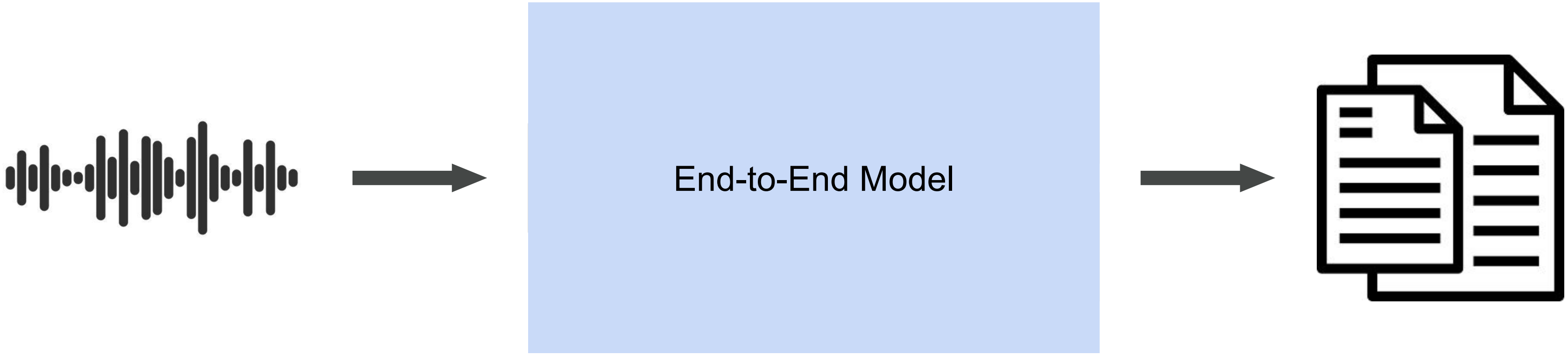
Table 3. CER and relative CER reduction of various evaluation sets

Evaluation set	Whisper model		
	large-v2	Model A	Model B
KsponSpeech eval set	13.95	9.44 (32.33)	**9.17** (34.26)
LibriSpeech test-clean	1.77	**1.19** (32.77)	1.33 (24.86)
LibriSpeech test-other	**2.86**	2.87 (–0.35)	3.39 (–18.53)

CER, character error rate.

Chang, J., & Nam, H. (2023) Exploring the feasibility of fine-tuning large-scale speech recognition models for domain-specific applications: A case study on Whisper model and KsponSpeech dataset. Phonetics Speech Sci. 2023;15(3):83-88. https://doi.org/10.13064/KSSS.2023.15.3.083

Speech-to-Text



Thank you!
Any Questions?

Special thanks to: my advisor Allison, committee members Matt & Marty,
Collaborators, FANCY lab, Luxlab (#gates214),
Family (for waking up @ 3am) & Friends