Habitat Learn

Addressing AI Issues In Automated Speech Recognition Models That Affect Transcription Accuracy

Accessing Higher Ground 2024

Why Do We Need To Know?

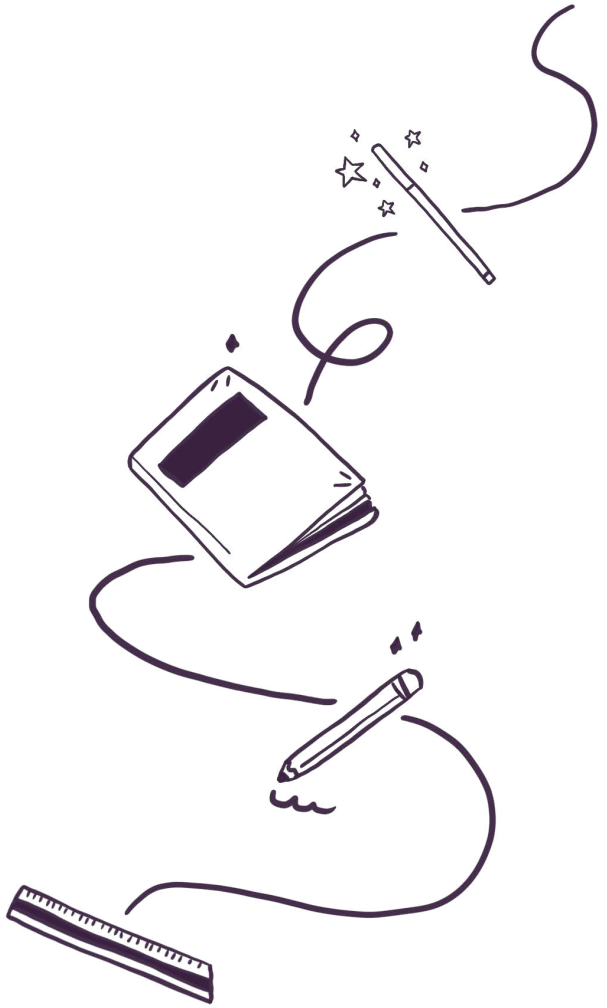
Automated Speech Recognition (ASR) has been around for a while

However, despite significant improvement in accuracy, ASR does not yet reach the required accuracy for many classroom situations

Domain specificities, technical terms and abbreviations, room acoustics, speaker speed, accent, age, gender and ethnicity can impact the accuracy – referred to as bias

If we understand the bias, can we use that bias to improve the accuracy of speech to text in our classrooms?

This talk is about a research project led by Habitat Learn into bias in ASR models



Who Are We – Our Credentials

Habitat Learn supports 100,000's of students at over 300 universities and colleges with barriers to learning arising from a learning difference, such as dyslexia, language difference or hearing loss.

We provide support through software and services:

Live captioning: Human captioners provide remote or in-class support for Deaf and hard of hearing students and some students with learning or language barriers.

Summary notes: Human note takers provide summary bullet point hierarchical notes for students with learning or language differences.

We have recorded over 1 million hours of lecture recordings and have used the datasets to create an unparalleled Large Language Model for education.

Messenger Pigeon can be used both for supporting our students with barriers to learning but also for every student as a productivity tool.

Messenger Pigeon

Messenger Pigeon can be used both for supporting our students with barriers to learning but also for every student as a productivity tool.

Recording: Messenger Pigeon records the lecture audio live or an uploaded video to produce an accurate transcript which is segmented and on playback highlights the relevant paragraph as the lecturer is talking.

Note taking: Users can take notes and make bookmarks of important passages in parallel with recording and/or on playback.

Captioning: Users can listen to or view a live recording and see live latency free captioning on screen.

Study Assistant: Creates question and answer sessions, keyword details, relevant videos to view, and study guide and tips on how to use their own content to reinforce learning.

Introduction To Research

Habitat Learn is leading a consortium supported by the UK government to identify bias in current Automated Speech Recognition Systems (ASRs).

The project is named DeepMyna.

The research will deliver an auditing tool which will enable us to evaluate different ASRs, the impact of different biases, and provide a benchmark for accuracy and bias.

The Consortium includes University of Southampton, Avanade and Microlink specialists in AI, Data consultancy and Assistive Technology.

Importance of accurate ASR in higher education

Accessibility requirements-captioning and transcripts

- Language and learning differences
- Deaf or hard of hearing
- Blind or low-vision
- Dysarthria or speech impairments

Universal Design for Learning

Impact of ASR Bias

Cultural marginalization
Minorities may feel excluded if ASR doesn't work well for them.

Economic implications
Poor transcription for certain accents may affect job opportunities in tech-heavy industries.

Accessibility
Individuals with disabilities or non-native speakers might face difficulties in using voice-enabled technologies.

Different Types of Bias

Accent and Dialect Bias: ASR models often have higher accuracy for standard accents than regional or global accents.

Gender Bias: ASR models generally perform better for male voices compared to female voices.

Ethnic and Racial Bias: Models may perform poorly on speech from ethnic minorities, contributing to disparities in usability.

Age Bias: May have higher error rates for children's or elderly people's speech.

Room Acoustics: May impact on quality of sound recording

Technology interference: May cause perturbations to the recording

Embedded Machine Learning Biases: Can impact accuracy through unmatched training

Causes of ASR Bias

Training Data Imbalance:

- ASR systems are typically trained on large datasets, but these datasets often underrepresent diverse accents, dialects, genders, and ethnic groups.

Model Architecture:

- Some neural network models are optimized for the majority group present in the data, leading to uneven performance across demographic groups.

Linguistic Complexity:

- Dialects and accents introduce variations in pronunciation, rhythm, and word usage that models may not capture effectively.

Socioeconomic Factors:

- Speakers from different socioeconomic backgrounds might use language differently, and these variations may be underrepresented in training data.

Consequences of ASR Bias

Exclusion
Certain groups might avoid using voice-activated services due to poor accuracy, leading to a digital divide.

Misinformation
Inaccurate transcriptions can lead to misunderstandings or incorrect outputs, especially in critical areas like healthcare or legal transcription.

Loss of Trust
Biased systems can undermine trust in technology, particularly among communities that are frequently misrepresented or marginalized.

Addressing Bias in ASR

Improving Data Diversity: Collect more representative datasets that include accents, dialects, genders, and ethnicities from around the world.

Bias Audits: Conduct regular audits of ASR systems to identify and mitigate performance gaps between different demographic groups.

Model Fine-tuning: Fine-tune models for specific groups or regions using targeted data.

User Feedback: Incorporate real-world user feedback to identify underperformance in diverse populations and improve model robustness.

How Do We Measure ASR Accuracy

Word Error Rate:

- Using a reference and recognised script and then count the number of substitutions (S), insertions (I), and deletion (D) errors divided by the total number of words (N)
- The lower the WER the better the ASR model

Other metrics:

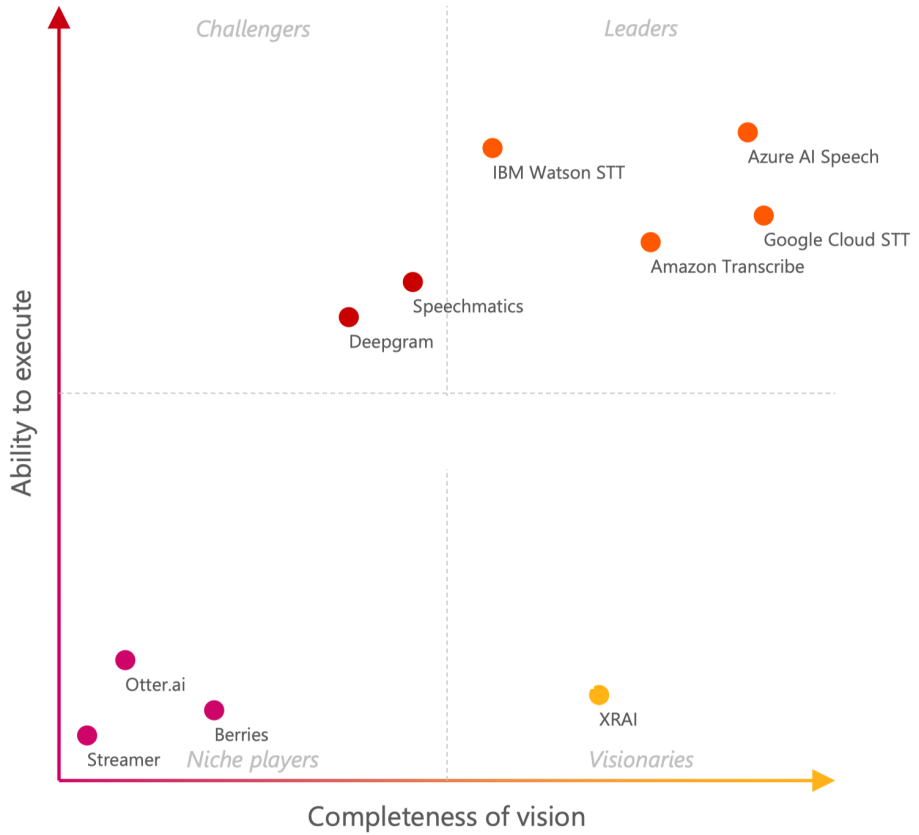
- Proper noun evaluation
- Choosing the right dataset – real world metrics
- Normaliser for domain specific subjects
- Contextual and semantic meaning

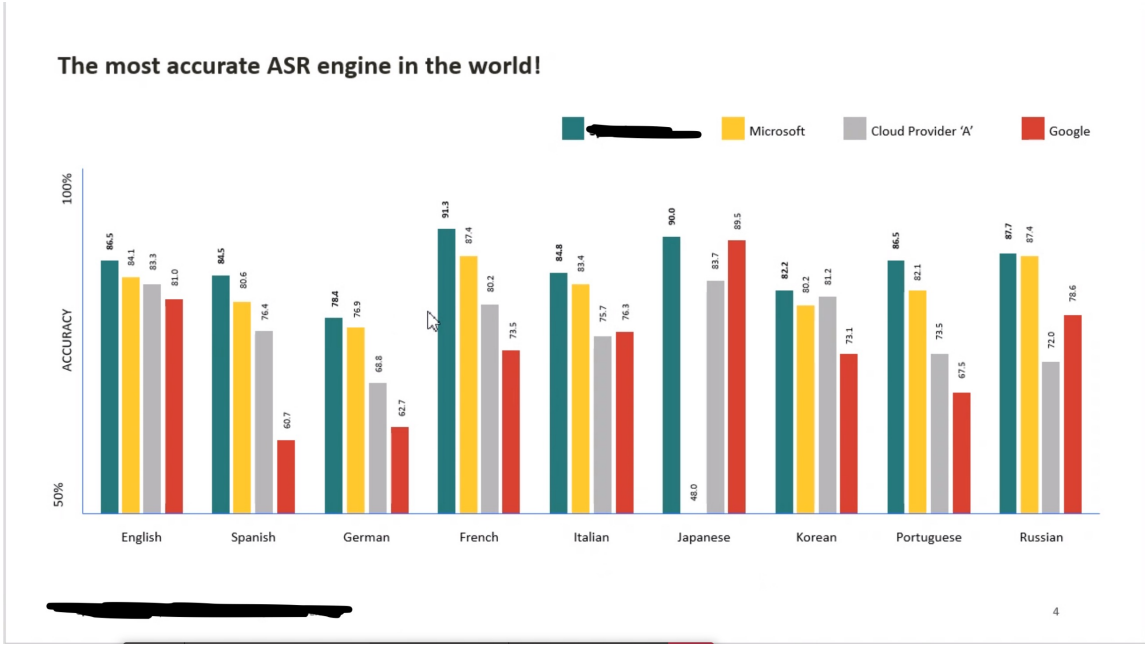
$$\mathrm{WER} = \frac{S + D + I}{N} \times 100$$

$$\mathrm{Accuracy} = 100 - \mathrm{WER}$$

The Major Providers of ASR Models

- Most large ASR models are a small proportion of the organisation business
- ASR engine are ’Off the shelf’ or ‘Out of the box’
- You can’t customise ASR with your own dataset
- Only care about general accuracy not accuracy on a certain bias
- Adding more training data irrelevant to the bias it will hurt the accuracy in that bias category, but boost the overall accuracy





Most Accurate ASR Model in the world!

- Each ASR company uses a different way to display their data to prove they are the “best”
- They use weighted average to boost the accuracy
- There is no transparency on bias – they just don’t know
- Labelling data to account for bias is human intensive and expensive
- Habitat Learn label data as part of their note taking and captioning activities and have over 1m lectures of labelled content

Accessing Higher Ground 2024

Weighted & Unweighted WER Using Open-Source Models

Example
- 10 sentences, 8 males, 2 females
- WER 90% for the male sentence, 40% for 2 female sentences
- Weighted average – sample size: (0.9*8+0.4*2)/10 = 80%
- Unweighted average: (90% + 40%)/2 = 65%

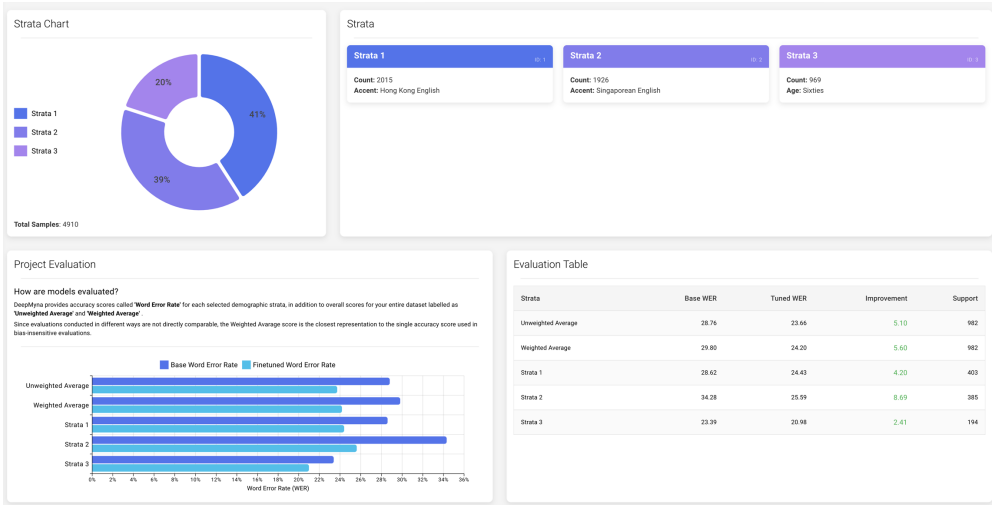
The problem
- Unweighted average accuracy can be much lower than weighed average
- 3rd Party ASRs only publish weighted average, but we don't know the sample size in each bias category
- Reality: we are likely to have a 50-50 chance to apply ASR to male-female (unweighted)
- So, the real accuracy is only 65% instead of 80%

Unbalanced training data leads to discrimination
- Adding more data to one bias category will boost the accuracy in that category only
- They add 10 more sentences to male and re-train it
- 18 male 98%, 2 female 40%
- Weighted average: 80% -> 92.2% (61% WER improvement)
- Unweighted average: 65%->69% (11% WER improvement)

Accessing Higher Ground 2024

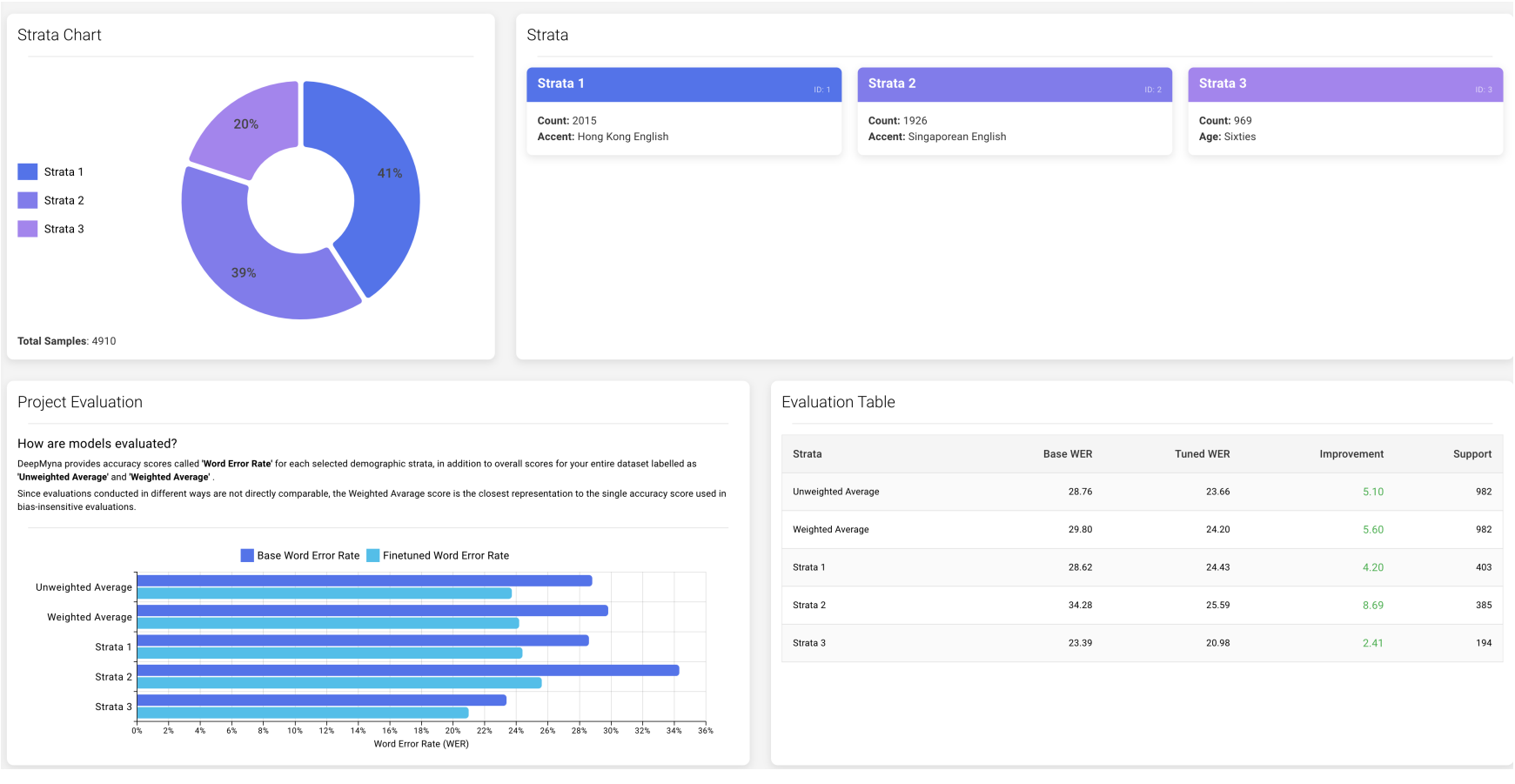
DeepMyna Evaluation Platform

- DeepMyna uses open-source ASR models to experiment with fine tuning bias in speech recognition
- We select sample sizes for test
- Divide the sample into two-one for test and the other for evaluation
- We then select different biases to empirically test whether we can improve the WER of a particular sample set



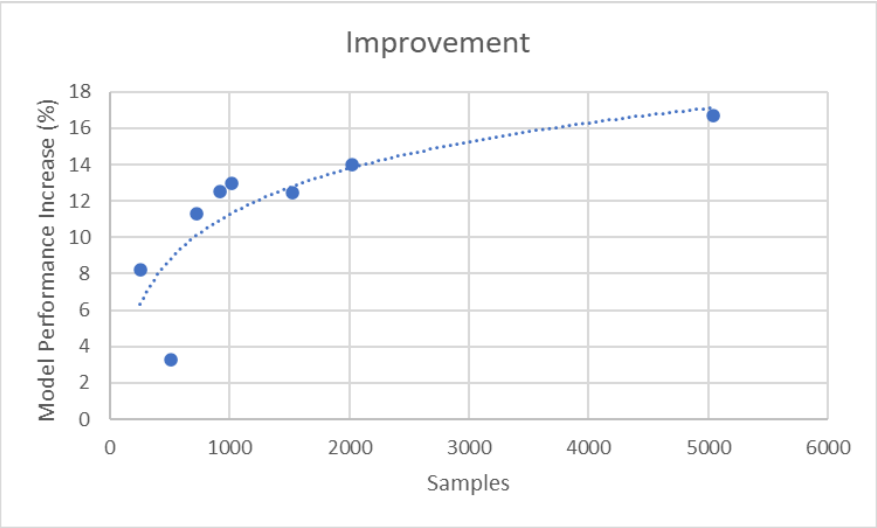
Accessing Higher Ground 2024

Live Demo of Evaluation Platform



Strata	Base WER	Tuned WER	Improvement	Support
Unweighted Average	28.76	23.66	5.10	982
Weighted Average	29.80	24.20	5.60	982
Strata 1	28.62	24.43	4.20	403
Strata 2	34.28	25.59	8.69	385
Strata 3	23.39	20.98	2.41	194

Example 1 Example 2



Fine-Tuning Efficiency

How much data is 'enough'

- Accuracy can improve by 50% with 15 hours training data
- Be prepared for decrease on minority category
- Trial and error
- High quality training data

Major biases to consider

- Accents
- Gender
- Age
- Domain knowledge
- Echo and environment noise

Balance the training dataset

- Which bias(es) you want to boost
- Which bias(es) you don't want to sacrifice
- Apply equal amount of fine-tuning data for those biases
- Accuracy will decrease for under-represented bias
- Exam the accuracy for weighted, unweighted average, and for each bias category (strata)

Benchmarking ASR Models DeepMyna Research Goals

Better transcription and captioning can improve learning outcomes

Benchmarked accuracy for all ASR solutions –choose which suits your user needs

Personalising user experience with datasets that match bias and domain specificity to improve accuracy

Cannot train core ASR provider models

Poor ASR transcripts or captions will yield poor Generative AI interactions

Innovate UK project - Avanade, Microlink, Affiniti AI, University of Southampton

References

1. Google ASR Study (2019)
- Source: A study conducted by Google researchers in 2019 found significant performance disparities in their ASR systems, particularly for African American Vernacular English (AAVE).
- Reference:
  - Koenecke, A., Nam, A., Lake, E., Nudell, J., Quartey, M., Mengesha, Z., Toups, C., & Norcie, G. (2020). Racial disparities in automated speech recognition. Proceedings of the National Academy of Sciences, 117(14), 7684-7689.
  - Available at: PNAS.org

2. Bias in Gender and Ethnic Recognition
- Source: Studies and research have demonstrated that ASR models show higher accuracy for male voices compared to female voices, as well as for white speakers compared to ethnic minorities.
- Reference:
  - Tatman, R. (2017). Gender and dialect bias in YouTube's automatic captions. Proceedings of the First ACL Workshop on Ethics in Natural Language Processing.
  - Available at: ACL Anthology

3. Mozilla Common Voice Project
- Source: Mozilla's open-source Common Voice project is a significant initiative aimed at reducing bias in ASR by collecting diverse speech data from around the world.
- Reference:
  - Mozilla Common Voice: commonvoice.mozilla.org

4. Challenges of Linguistic Diversity in ASR
- Source: Several academic and industry research papers highlight how ASR models struggle with linguistic diversity, especially when dealing with non-standard accents or dialects.
- Reference:
  - Huang, X., Baker, J., & Reddy, R. (2014). A historical perspective of speech recognition. Communications of the ACM, 57(1), 94-103.
  - Available at: ACM.org

References -continued

5. Multilingual and Cross-Dialect ASR Systems
- Source: Research shows that multilingual ASR systems using transfer learning or multi-task learning are increasingly being developed to mitigate bias across languages and dialects.
- Reference:
  - Gales, M. J. F., & Young, S. J. (2008). The application of hidden Markov models in speech recognition. Foundations and Trends® in Signal Processing, 1(3), 195-304.
  - Available at: Now Publishers

6. Ethical AI Frameworks in ASR Development
- Source: Ethical guidelines for reducing bias in AI and ASR systems have been proposed by both academia and industry, emphasizing the need for inclusivity and transparency.
- Reference:
  - Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency.

7. Accuracy of Speech to Text captioning in perfect acoustic environments
- Source: Accuracy of speech to text for speakers with native English accents and perfect acoustic environments
- Reference:
  - Millet,P (2021). Accuracy of Speech-to-Text Captioning for Students Who are Deaf or Hard of Hearing. Journal of Educational, Pediatric & (Re)Habilitative Audiology Vol.25 2021-22

8. Common sense approach to speech recognition
- Source: Accuracy of speech to text for speakers with native English accents and perfect acoustic environments
- Reference:
  - Henry Lieberman, Alexander Faaborg, Waseem Daher, José Espinosa
  - How to Wreck a Nice Beach You Sing Calm Incense. Web.media.mit.edu

Habitat Learn

Contact Us
Jeremy.Brassington@habitatlearn.com (+447785 225600)
Daniel.Goerz@habitatlearn.com (+15126802193)

www.habitatlearn.com