Problem Set #1

synchrony FINANCIAL

Equitability in Voice Assistants

Aparna Ayyah & Matt Ho (Team 77)

Methodology

1. **PROBLEM CONTEXT**
   Current Voice Assistant limitations

2. **DATASETS OVERVIEW**
   Indic TTS + CommonVoice

3. **CUSTOM MODEL PERFORMANCE**
   RNN Model + Spectrograms

4. **TEXT UNDERSTANDING**
   N-Gram Model Improvements

5. **CONCLUSIONS**
   Summary of Improvements

Inequity in Voice Assistants

Dataset

Historically drawn from standard, “White” dialects

Prediction Error

Word error rate up to 2x higher for accented speakers vs White speakers

Positive Feedback Loop

Skews future models because of current error

Koenecke, Allison, et al. “Racial Disparities in Automated Speech Recognition.”

Key Underserved Factors

Accented Speech

Those with accents may pronounce or phrase commands differently

Speech Impairment

Voice Assistants have a long way to go with accessibility

Children

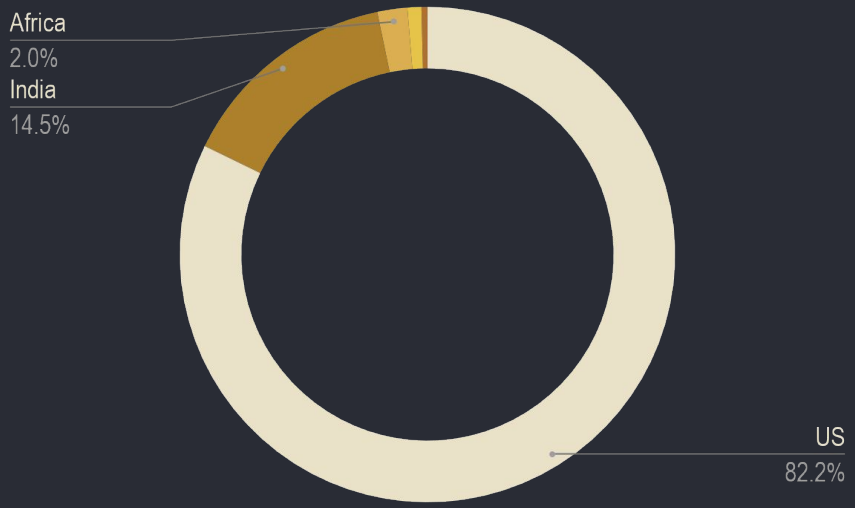
Children interact with and arrange words differently than adults

Scanlon, Dr. Patricia. "Voice Assistants Don't Work for Kids: The Problem with Speech Recognition in the Classroom."

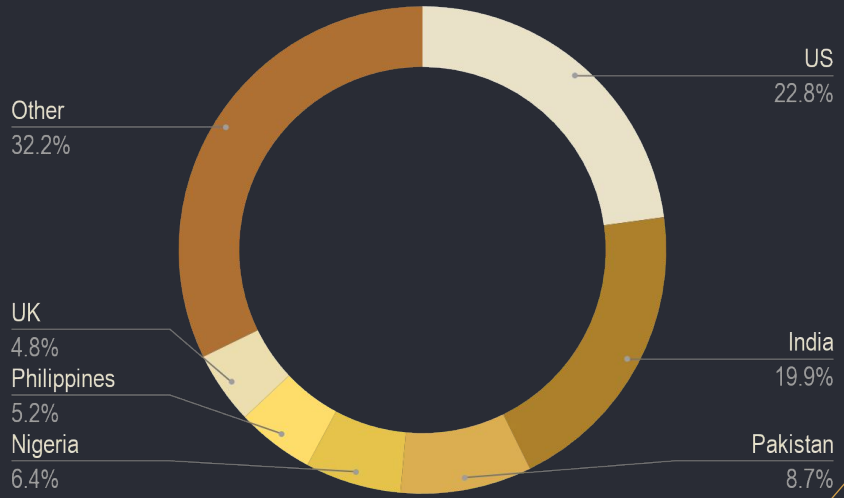
Masina, Fabio, et al. "Investigating the Accessibility of Voice Assistants With Impaired Users: Mixed Methods Study."

Data Composition

Composition of CommonVoice Dataset

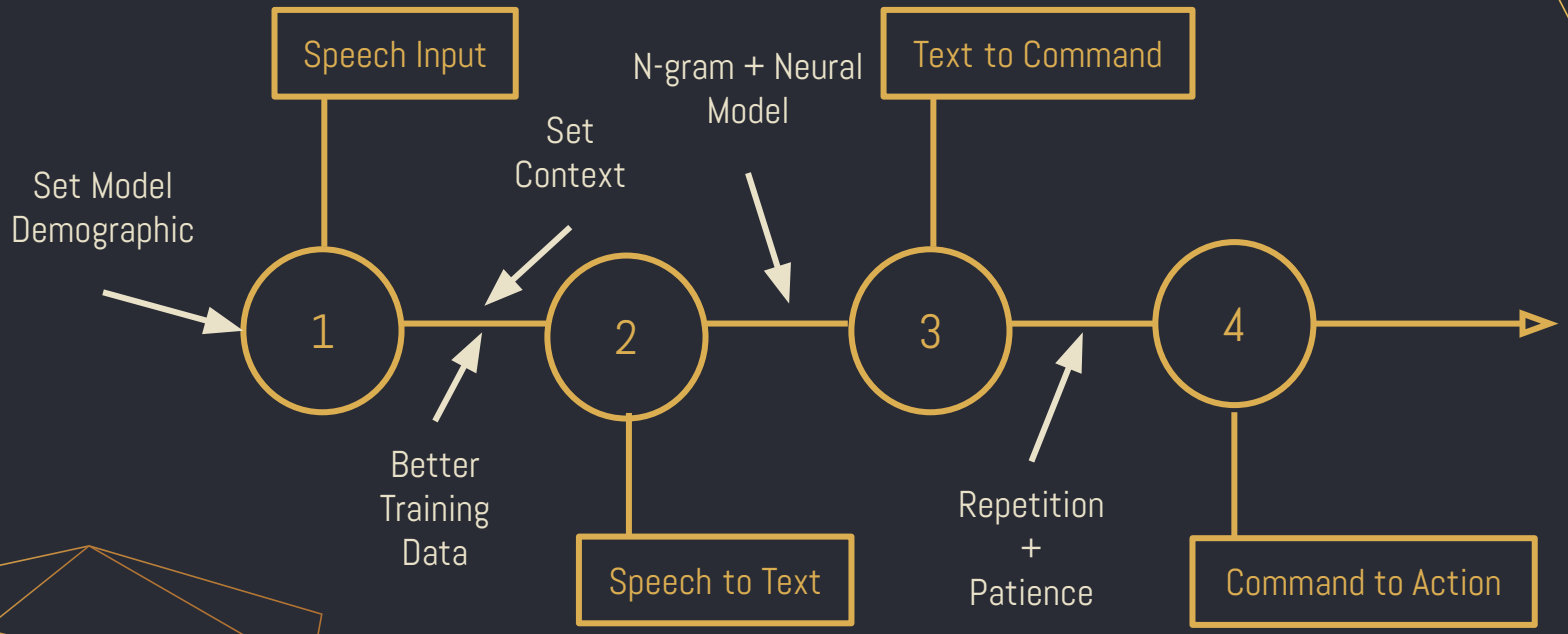


English Speakers Worldwide



The Training Data is not representative of Users.

Voice Assistant Sequence



Improvements in Speech to Text

Datasets

Indic TTS

Open Source Speech Dataset:

- Created by Indian Government
- Has extensive audio of Indian accents + phrasing
- Recorded by researchers
- Used for our Custom Model

CommonVoice

Open Source Speech Dataset:

- Created by Mozilla
- Similar to the (proprietary) datasets used by Amazon, Google, etc.
- User and Volunteer Generated
- Used by DeepSpeech (Voice Recognition Engine)

Anusha Prakash, Anju Leela Thomas, S. Umesh and Hema A. Murthy, "Building Multilingual End-to-End Speech Synthesisers for Indian Languages".

Custom Model Components

.wav files

Spectrogram

Allows for transformation of audio files into data points

Normalization

Normalize pixels of spectrogram

Recurrent Neural Network

Better for sequential data (Long Term Short Memory)

TimeDistributed Layer

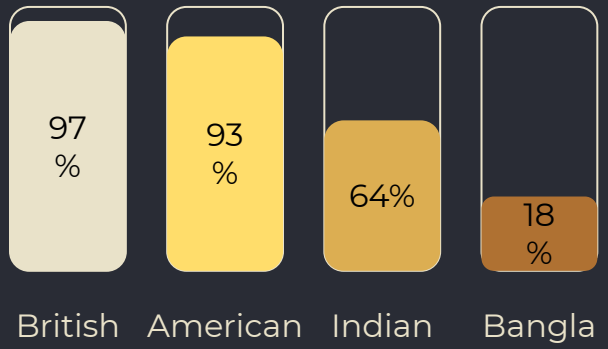
Layer applied to each input (useful for sequential data)

V, Anand P. "Indian Accent Speech Recognition."

Comparison of Models

	Generic DeepSpeech	Trained Custom Model
Training Data	Standard	Indian Accents
Epochs	20	20
Word Error Rate	0.44	**0.16**
Word Accuracy	0.56	**0.84**

Prediction Confidence of DeepSpeech



V, Anand P. "Indian Accent Speech Recognition."

K-Fold Cross Validation

V, Anand P. "Indian Accent Speech Recognition."

Improvements in Text to Action

N-Gram Model

N-Gram Method

- Used by conventional models
- Calculate probability of the next word given the past n-1 words
- N = 4 (for large scale, working models)

Advantages

- Uses some context in prediction (N amount)

Limitations

- Misses longer range dependencies

Modified N-Gram Model

Step 1
Run conventional N-gram method

→

Step 2
Refine results using Neural Model

→

Step 3
Use Noise Contrastive Estimation to refine results

↓

Step 4
Use Quantization methods to improve computing time

←

Result
- Reduces Error Rate by 6.2%
- Increases computing time by just milliseconds

Raju, Anirudh. "How to Make Neural Language Models Practical for Speech Recognition."

Additional Improvements

Setting Context

- Classifying by Intent:
  - Saying "Temperature" for temperature related commands
  - Adds to time taken, but increases accuracy

Repetition + Learning

- Ask again if unsure:
  - "Sorry, could you repeat that?"
- Track frequent errors

L. Lugosch, M. Ravanelli, P. Ignoto, V. Tomar, and Y. Bengio, "Speech Model Pre-training for End-to-End Spoken Language Understanding"

Conclusions

Speech to Text

- Currently training datasets are not robust enough
  - Natural Variations (like accent) can and will occur
- Mirroring population composition with training data improves model performance
  - Users can input preferred model (“select dialect”) to improve voice assistant performance

Text to Action

- Text to Action understanding can be improved
  - Understand long range context
- Allow User to set “context”
  - Helps reduce misunderstandings in word choice

Thank You!

synchrony

"Compared to 'traditional' forms of discrimination, automated discrimination is more abstract and unintuitive, subtle, intangible, and difficult to detect"

—AI Expert Sandra Wachter