Speech recognition for multi-channel communication networks

Dr. Tejas Shastry
Chief Data Scientist, GreenKey Technologies
tejas@greenkeytech.com
Voice assistants are taking over homes. Why aren’t they taking over businesses?
Business conversations are noisy and distributed.
Home assistants can send high quality audio to the cloud

Audio quality can be maintained if you always control the microphone
Business conversations are distributed across often noisy environments

Disparate audio quality leads to poor speech recognition accuracy
How does modern speech recognition work?

Training

- Audio Files
- Text

Inference
How does modern speech recognition work?

Training

Audio Files → Feature Transformation → Numbers

Text → Feature Transformation → Numbers

Inference
How does modern speech recognition work?

**Training**
- Audio Files
- Feature Transformation
- Text
- Numbers
- Target Output

**Inference**
- Neural Network Model
- Output
How does modern speech recognition work?

Training

Audio Files → Feature Transformation → Numbers → Neural Network Model

Text → Feature Transformation → Numbers

Input → Neural Network Model → Output

Target Output

Compare the outputs, update the model

Inference
How does modern speech recognition work?

**Training**
- Audio Files
  - Text
  - Feature Transformation
  - Numbers
    - Input
    - Target Output
    - Neural Network Model
      - Output
      - Compare the outputs, update the model

**Inference**
- Audio
  - Feature Transformation
  - Numbers
    - Input
    - Neural Network Model
      - Output
      - Transcription
How does modern speech recognition work?

Speech recognition models expect audio features to be similar to those they were trained on.
Speech recognition accuracy depends on audio quality

- Telephony Streams
  - 4 kHz – 13 kHz
- SIP Streams
  - 16 kHz – 32 kHz
- WebRTC Streams
  - 40+ kHz

Audio Score:
- 4 kHz – 13 kHz: 55%
- 16 kHz – 32 kHz: 80%
- 40+ kHz: 92%
- 40+ kHz: 95%
What “audio quality” features affect speech recognition accuracy?

• Sampling rate

• Signal to noise ratio

• Representation of “human voice” frequencies

• Clipping / distortion

• Mean amplitude / loudness
Train various speech recognition models on varied audio quality to improve overall accuracy
Train various speech recognition models on varied audio quality to improve overall accuracy

Performance of Individual Models vs. Composite(s)

Word error rate (Lower is better)

- Model A: 52%
- Model B: 32%
- Model C: 29.7%
- Model D: 23.2%
- Model B+D: 20.8%
- Model C+D: 20.1%
- Model A+B+C: 19.2%
- All 4 models: 15.3%
Train various speech recognition models on varied audio quality to improve overall accuracy

https://www.greenkeytech.com/video/scribe/Noisy_Quotes.mp4
Train various speech recognition models on varied audio quality to improve overall accuracy
Business conversations are **noisy** and **distributed**.

Speech recognition needs to be trained to understand each unique audio environment.