Computational modeling of speech recognition offers a promising tool to study human speech perception.

### INTRO
Humans extract acoustic features from audio to identify phonemes to decode words. We can simulate this process with machine learning (neural networks).

Pytorch contains off-the-shelf neural networks trained to do tasks like speech recognition.

**Backbone:** neural network that contains the mapping from audio to phoneme.

**Decoder:** algorithm that decides the phonemes based on the audio.

Automatic Speech Recognition (ASR): neural networks trained to convert audio into text.

### METHODS
(adapted from Hira, M., 2024)

**Get Audio**

**Clean all files (remove noise, extra words)**

**Match sampling rate to PyTorch requirements**

**Load PyTorch**

**Load backbone & decoder**

**Run ASR**

**Calculate Word Error Rate**

**Word Error Rate (Levenshtein Distance):** difference between two texts/strings.

### RESULTS

**Simulation of Accent Perception**

**Simulation of Accent Training**

**No Training**  |  Single Accent  |  Multi-Accent
---|---|---
American Accented English | British Accented English | Indian Accented English

### METHODS

- Greedy: Pick the most likely option at each decision
- Beam Search: Pick the most likely option over n words
- Substitutions

### FUTURE DIRECTIONS
- Validate automatic speech recognition models as tools for human speech perception.
- Simulate learning/training studies via fine-tuning (i.e. fine-tune an English ASR model with Spanish to simulate Spanish language learning).

### REFERENCES


### DETAILS

**Backbone:** wav2vec2

**Types of Decoders**

1. Greedy: Pick the most likely option at each decision
2. Beam Search: Pick the most likely option over n words

**Word Error Rate**

Types of Changes:

1. Addition
2. Insertions
3. Substitutions

Response | Answer | Word Error Rate
---|---|---
LUCK | BLACK | 100%
RECIPT | RECEIPT | 0%
OUR | OUR | 100%

This measure is imperfect; this mistake would count as 2 errors instead of 1 (swap E and I)

Scoring homophones causes issues (1/3 errors and 1/4 errors)

**Likelihood of Phoneme**

**Decoder**

**Types of Decoders (non-inclusive):**

- Greedy
- Beam Search

**Accuracy:**

- 0.1
- 0.2
- 0.3
- 0.4
- 0.5
- 0.6
- 0.7
- 0.8
- 0.9
- 1.0

**Speech in Noise**

- No Foreign Accent
- Single Foreign Accent
- Multi-Accent

**Simulation of Accent Perception**

American Accented English  |  British Accented English  |  Indian Accented English
---|---|---

**Simulation of Accent Training**

No Training  |  Single Accent  |  Multi-Accent
---|---|---
American Accented English | British Accented English | Indian Accented English

**Word Error Rate:**

- 0
- 1
- 2
- 3
- 4


**REFERENCES**


