BABBLE BUDDY
Online Speech Therapy - Phoneme Recognition

MIDS Capstone Project by
Tyrnan Prasad, Derek Lee, Elena Xie
Agenda

1. Project Overview
   Objective & Mission

2. Minimum Viable Product
   MVP Process Flow
   MVP Demo

3. Technical Approach
   Model development process
   Pretrained models
   Model evaluation
   User feedback generation

4. Closing Remarks
   Summary of the project
   Roadmaps
Project Overview

Our objective and mission
Ideation & Objective

Problems we are solving
Balancing the growing demand with current supply shortfall in Speech-Language Pathologists (SLPs) shortage

Our objective
Introducing a digital tool for online speech therapy, ensuring accessibility and effectiveness in reaching children with speech disorders.

Mission
Help every child speak with confidence
Product Usage

Support SLPs
- Offers real time feedback on speech patterns, facilitating more efficient therapy sessions.

Support Parents
- Provide access to resources that can be conducted at home
- Enhancing continuity of care outside of traditional therapy session

Self-Management
- Encourages independent learnings for teenagers
3 - 17 years old
Targeted Users

6 Million
Potential Market Size

Easy and Reliable
Is our key advantage
Minimum Viable Product

Let’s look at the product
MVP Flow

Front End
1. Sentences appear on-screen. User can start recording
2. User clicks the 'Ready' button on the front end to submit a request for feedback

Back End
3. Model process user recordings, generating sequence of phonemes
4. Feedback generation by comparing model prediction vs. correct phonemes

5. Visualizing feedback for user interpretation
MVP DEMO

Welcome To Online Speech Therapy!
IT'S NICE TO MEET YOU

OUR PRODUCT
Our goal is to provide targeted support to speech therapy patients, particularly children struggling with articulation. Our product promises an effective therapy experience for our young patients, empowering them to overcome speech challenges with confidence.

Key Values
- Efficient accessibility
- Child-friendly usage
- High educational standard

Product Design
Our tool is easy to use. Simply record your voice by reading the sentence generated on the screen. Click ready to get feedback!

Visit Our App

Information Security
No user information is stored. Your privacy is our top priority.
Technical Approach
Automatic Speech Recognition (ASR)

**Speech Audio** → **Split into Frames** → **Extract Features** → **Classify Frames** → **Predict Sentence**

- **Acoustic Model**
- **Language Model**

**Traditional ASR**: Predict word sequence given detected phonemes

**Babble Buddy**: Classify pronunciation errors given detected phonemes and known word sequence
Project Dataset

**Original dataset**
(DARPA-TIMIT Acoustic-Phonic Continuous Speech Corpus)

**Custom dataset**
(derived from DARPA-TIMIT)

**About it**
→ Sound files paired with phonetic transcriptions with timestamps
→ Adult American English speakers grouped by accent from 8 regional

**Things we did**
→ Split DARPA-TIMIT data into discrete phonemes
→ Recombined phonemes randomly
→ Enforced class balance
→ Incorporated data augmentation (noise adding, pitch shift)

<table>
<thead>
<tr>
<th></th>
<th>Utterances</th>
<th>Minutes of speech</th>
</tr>
</thead>
<tbody>
<tr>
<td>Train</td>
<td>4,620</td>
<td>137 mins</td>
</tr>
<tr>
<td>Test</td>
<td>1,680</td>
<td>34 mins</td>
</tr>
</tbody>
</table>
Developing a Model for Babble Buddy

The Options:

<table>
<thead>
<tr>
<th>Custom Model</th>
<th>Pre-trained Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fit-for-purpose (phoneme recognition)</td>
<td>Trained for word recognition</td>
</tr>
<tr>
<td>Lightweight (vocab size ~60 phonemes)</td>
<td>Large (vocab size ~100,000 words)</td>
</tr>
<tr>
<td>Flexible</td>
<td>Fixed architecture, may not allow fine-tuning</td>
</tr>
<tr>
<td>Large development effort required</td>
<td>Minimal development effort</td>
</tr>
</tbody>
</table>
# Encoder-Decoder Model Experiments & Results

<table>
<thead>
<tr>
<th>Mel bands</th>
<th>Spectrogram Width</th>
<th>Dataset</th>
<th>n_fft</th>
<th>Length of Feature Vector / LSTM Units</th>
<th>Learning Curve (Loss)</th>
<th>Validation Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>128</td>
<td>281</td>
<td>Custom</td>
<td>2048</td>
<td>256</td>
<td></td>
<td>0.453</td>
</tr>
<tr>
<td>64</td>
<td>281</td>
<td>Custom</td>
<td>2048</td>
<td>256</td>
<td></td>
<td>0.532</td>
</tr>
<tr>
<td>32</td>
<td>281</td>
<td>Custom</td>
<td>2048</td>
<td>256</td>
<td></td>
<td>0.460</td>
</tr>
<tr>
<td>64</td>
<td>374</td>
<td>Custom</td>
<td>2048</td>
<td>256</td>
<td></td>
<td>0.432</td>
</tr>
<tr>
<td>64</td>
<td>200</td>
<td>Custom</td>
<td>2048</td>
<td>256</td>
<td></td>
<td>0.423</td>
</tr>
<tr>
<td>64</td>
<td>256</td>
<td>DARPA-TIMIT</td>
<td>2048</td>
<td>256</td>
<td></td>
<td>0.701</td>
</tr>
<tr>
<td>64</td>
<td>256</td>
<td>DARPA-TIMIT</td>
<td>256</td>
<td>256</td>
<td></td>
<td>0.697</td>
</tr>
<tr>
<td>64</td>
<td>256</td>
<td>DARPA-TIMIT</td>
<td>256</td>
<td>128</td>
<td></td>
<td>0.700</td>
</tr>
</tbody>
</table>
Custom Model Challenges

**Dataset Size**
- DARPA TIMIT dataset is limited in size (4,620 training examples), which limits the model’s ability to capture sufficient variability in the data.
- Larger datasets not freely available

**Data Augmentation**
- Time and frequency masking are detrimental to model’s ability to learn patterns, was removed

**Padding**
- Accuracy scores included prediction on the “pad” token, leading to inflated accuracy (actual accuracy about 50% lower than reported if pad tokens accounted for)
Pretrained Model: Allosaurus

**Allosaurus**

- **Universal Phone Recognizer**: Pre-trained on 2+ million utterance from 14 languages
- **Architecture**: Similar to transitional ASR systems, tailored for universal application
- **Feature Extraction**: Waveform $\rightarrow$ Open-Source Feature Extractor $\rightarrow$ 40-dimensional MFCCs
- **Encoder**: MFCCs $\rightarrow$ 6-layer bidirectional LSTM $\rightarrow$ Universal phone prediction layer $\rightarrow$ Allophone prediction layer $\rightarrow$ Phoneme

Allosaurus reference:
https://github.com/xinjli/allosaurus
# Model Evaluation Metrics - BLEU

<table>
<thead>
<tr>
<th>Baseline Dictionary Model</th>
<th>Allosaurus Model</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>0.456</strong></td>
<td><strong>0.473</strong></td>
</tr>
</tbody>
</table>

Given a prompt sentence, assume that the correct phonetic transcription of the response will be the dictionary correct response.

Reference tables used in BLEU:
- Mapping IPA phonetic symbol to Alphabets phonemes (i.e. translation of diphthongs)
- Flattening of the phonemes (i.e.: tcl to t)
User Feedback

- Minimum Edit Distance algorithm assigns error to particular sounds
- Flattening of sounds into larger categories
- Sounds correspond to words
- SLP defined error categories are individually labeled and returned

![Phonetic Symbols Table]

<table>
<thead>
<tr>
<th>Phonetic Symbols</th>
<th>Sounds</th>
<th>Photos</th>
<th>Drawings</th>
</tr>
</thead>
<tbody>
<tr>
<td>æ, ei</td>
<td>at, and,</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>ate</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0, 3, ɑ, r</td>
<td>look, bird,</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>supply, red</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ɑ, ʌ, ai</td>
<td>dog, cut,</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>ice</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ɛ, ɪ</td>
<td>end, it</td>
<td></td>
<td></td>
</tr>
<tr>
<td>i, j, s, ʃ</td>
<td>eat, yes,</td>
<td></td>
<td></td>
</tr>
<tr>
<td>z, ʒ</td>
<td>so, show,</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>zoo, vision</td>
<td></td>
<td></td>
</tr>
<tr>
<td>u, ʊʊ, w</td>
<td>you, no,</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>were</td>
<td></td>
<td></td>
</tr>
<tr>
<td>b, m, p</td>
<td>but, man,</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>pet</td>
<td></td>
<td></td>
</tr>
<tr>
<td>tf, t</td>
<td>chat, tea</td>
<td></td>
<td></td>
</tr>
<tr>
<td>d, g, ʤ, k, n, ŋ</td>
<td>dim, go,</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>jog, king,</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>new, sing</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ɔ, l, ɶ</td>
<td>the, lie,</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>think</td>
<td></td>
<td></td>
</tr>
<tr>
<td>f, v</td>
<td>fat, view</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Closing Remarks
Roadmaps

Key Learnings

- Phonemized Data produces patterned errors in trained models
- Context is relevant, rearrangement of sounds not viable strategy

Future Improvements

If more time and budget allowed:
- Dataset: Manually labelled dataset and a larger dataset
- Engagement Enhancement: gamify the product to engage targeted audience
Help Every Child Speak with Confidence
Appendix 1: Audio Data 101

What is a phoneme?
- The smallest unit of speech sound distinguishing one word element from another.
- Model output

What is a spectrogram?
- A visual representation of the spectrum of frequencies in a sound.
- Model input

What is a waveform?
- Displays changes in a signal's amplitude over time.
- Not used in model but will display on MVP when user is recording
Appendix 2: Encoder-Decoder Model

Example: the encoder-decoder model used for inference using the word “cat”

Inputs: previously generated phonemes

Outputs: predicted sequence of phonemes

Vocab Size = 64

Shape = (n_mels, width, 1)

Example: the encoder-decoder model used for inference using the word “cat”

Shape = (n_mels, width, 1)