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

1

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

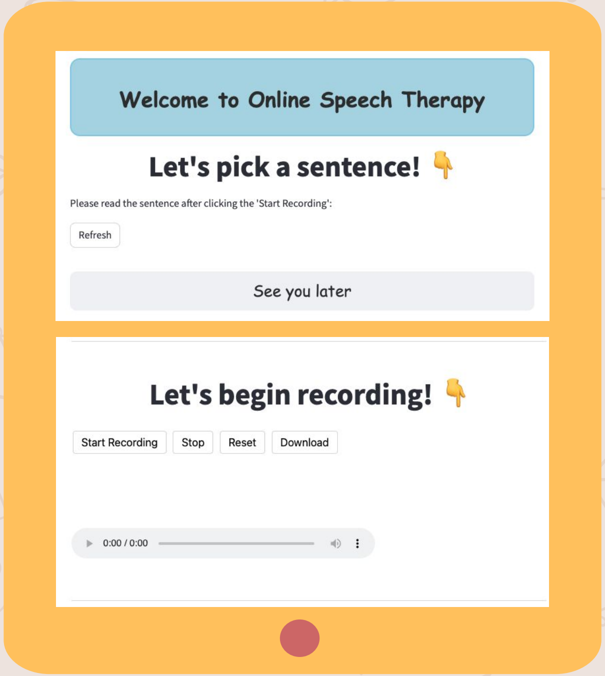
Our objective

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

2

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

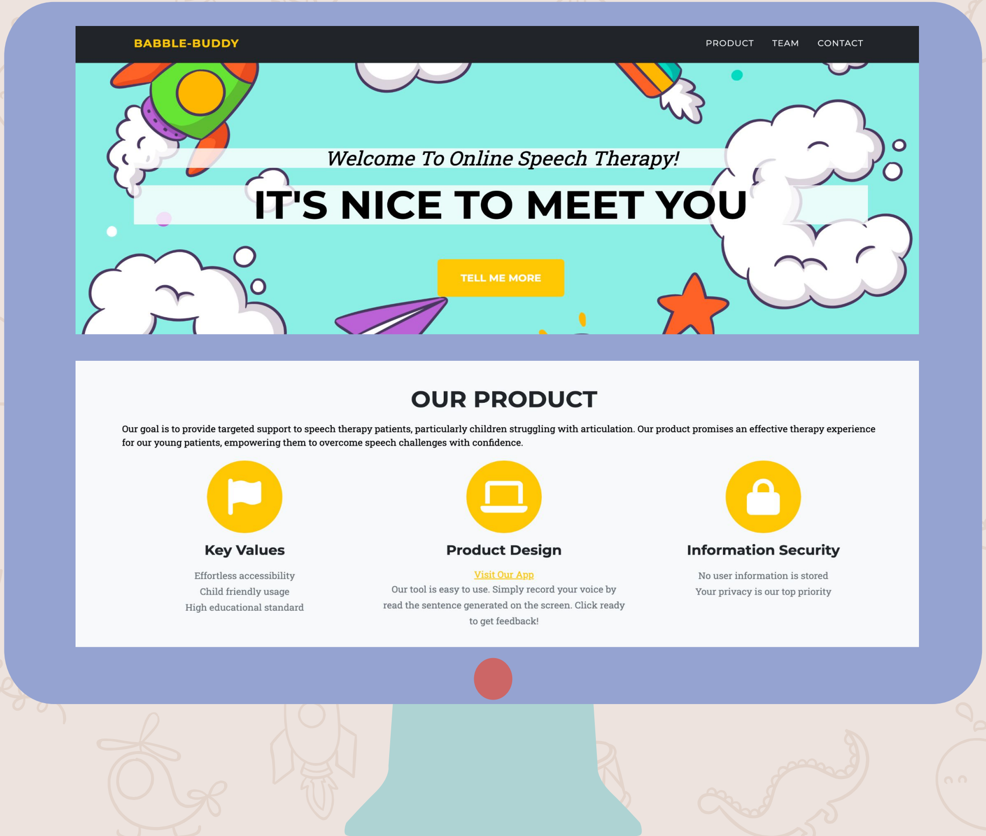
5 Visualizing feedback for user interpretation

Back End

Model process user recordings, generating sequence of phonemes 3

Feedback generation by comparing model prediction vs. correct phonemes 4

MVP DEMO



Visit our website here

3

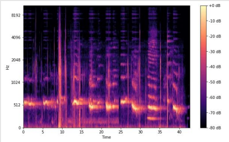
Technical Approach

Automatic Speech Recognition (ASR)

Speech Audio → Split into Frames → (Audio Frames) → Extract Features → (Feature Frames) → Classify Frames → (Phoneme Sequence) → Predict Sentence → (Word Sequence)

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-Phonetic Continuous Speech Corpus)

About it

→ Sound files paired with phonetic transcriptions with timestamps
→ Adult american english speakers grouped by accent from 8 regional

	Utterances	Minutes of speech
Train	4,620	137 mins
Test	1,680	34 mins

Custom dataset
(derived from DARPA-TIMIT)

Things we did

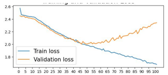
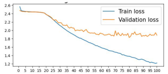
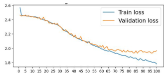
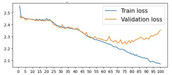
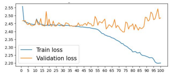
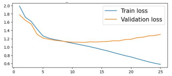
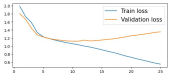
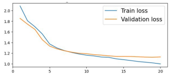
→ Split DARPA-TIMIT data into discrete phonemes
→ Recombined phonemes randomly
→ Enforced class balance
→ Incorporated data augmentation (noise adding, pitch shift)

Developing a Model for Babble Buddy

The Options:

Custom Model	Pre-trained Model
Fit-for-purpose (phoneme recognition)	Trained for word recognition
Lightweight (vocab size ~60 phonemes)	Large (vocab size ~100,000 words)
Flexible	Fixed architecture, may not allow fine-tuning
Large development effort required	Minimal development effort

Encoder-Decoder Model Experiments & Results

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

Custom Model Challenges

Limited size

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

Not improving model

Data Augmentation

→ Time and frequency masking are detrimental to model's ability to learning patterns, was removed

Inflating accuracy

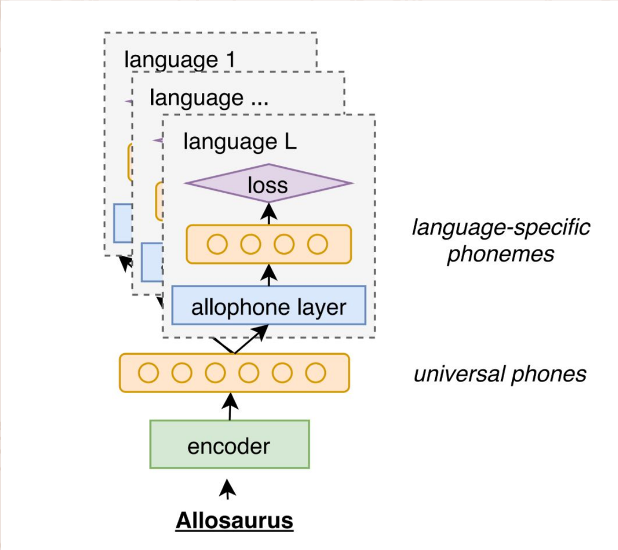
Padding

→ 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 → Open-Source Feature Extractor → 40-dimensional MFCCs
- **Encoder:** MFCCs → 6-layer bidirectional LSTM → Universal phone prediction layer → Allophone prediction layer → Phoneme



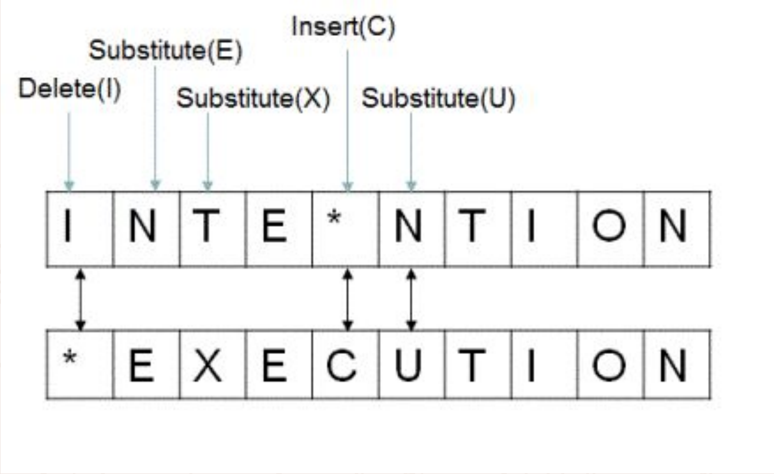
Allosaurus reference:
https://arxiv.org/pdf/2002.11800.pdf
https://github.com/xinjli/allosaurus

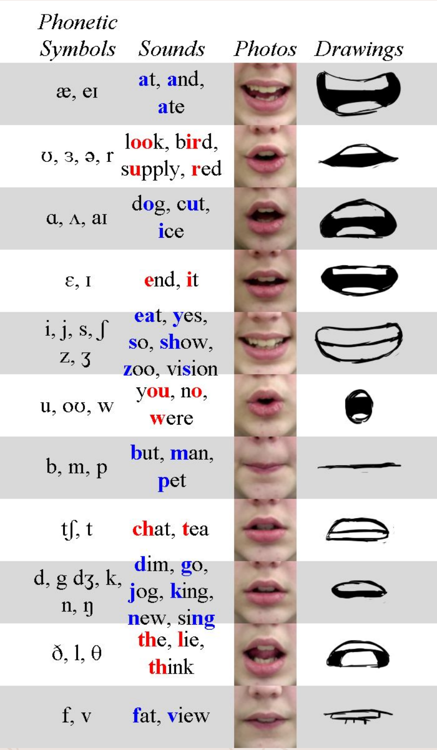
Model Evaluation Metrics - BLEU

Baseline Dictionary Model	Allosaurus Model
0.456	0.473
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 (ie. 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	Sounds	Photos	Drawings
æ, eɪ	at, and, ate		
ʊ, ɜ, ə, r	look, bird, supply, red		
ɑ, ʌ, aɪ	dog, cut, ice		
ɛ, ɪ	end, it		
i, j, s, ʃ z, ʒ	eat, yes, so, show, zoo, vision		
u, oʊ, w	you, no, were		
b, m, p	but, man, pet		
tʃ, t	chat, tea		
d, g dʒ, k, n, ŋ	dim, go, jog, king, new, sing		
ð, l, θ	the, lie, think		
f, v	fat, view		

4

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

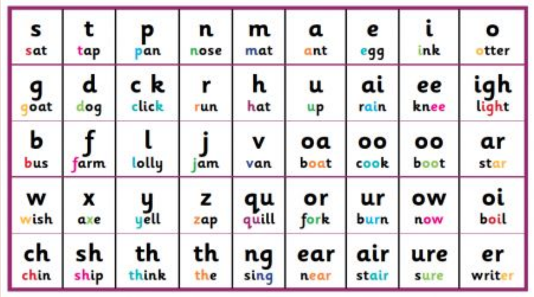
5

Appendix

Appendix 1: Audio Data 101

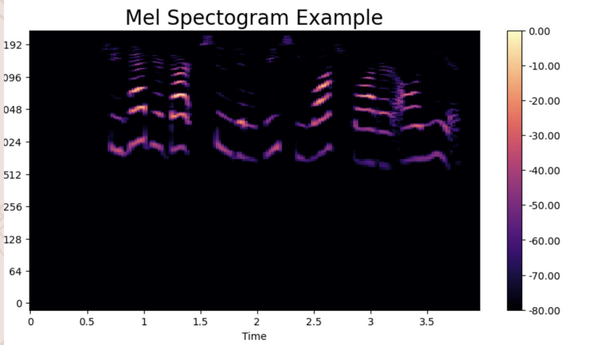
What is a phoneme?

- The smallest unit of speech sound distinguishing one word element from another.
- Model output



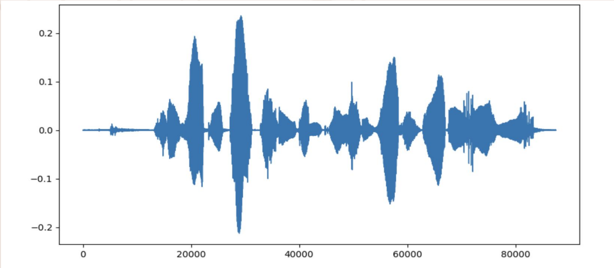
What is a spectrogram?

- A visual representation of the spectrum of frequencies in a sound.
- Model input

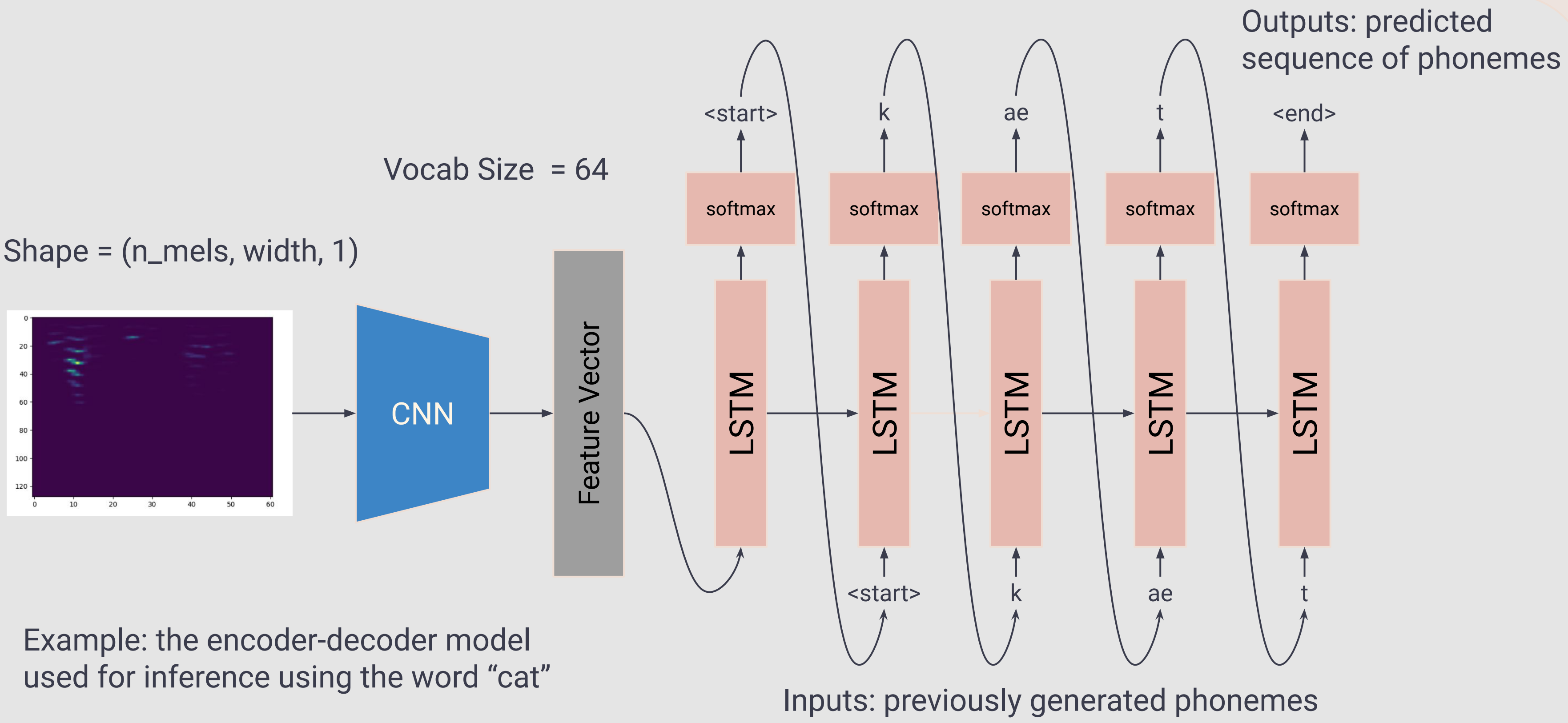


What is a waveform?

- 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”