# Speech-to-Text Engine

**E533 – Deep Learning Systems**

*Submitted by:* Ankit Mathur & Saurabh Mathur

# Main Dataset


- 200K+ audio samples
- 5K+ speakers
- 200+ hours
- 9 Age groups, 2 Genders, 6 Accents

# Accent Classification: Dataset

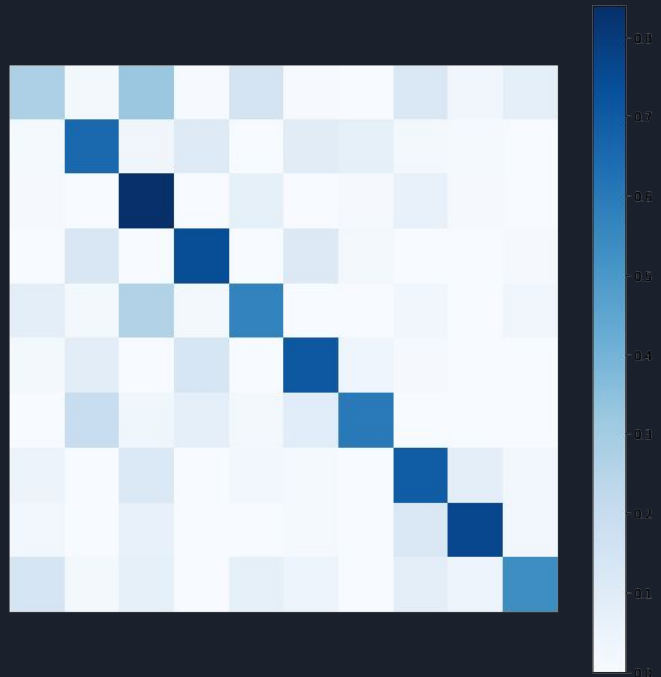

- 1K audio samples
- 1K speakers
- 5+ hours
- 10 combinations of Gender x Accent

# Accent Classification: Model

```
Net(
  (lstm): LSTM(513, 128, num_layers=2, batch_first=True)
  (linear): Linear(in_features=256, out_features=10, bias=True)
)
```

- Predicts: gender - accent
- Optimizer: Adam with weight decay
- Loss: Class-weighted cross-entropy

# Accent Classification: Results



**Confusion Matrix**

# Speech to Text: Dataset

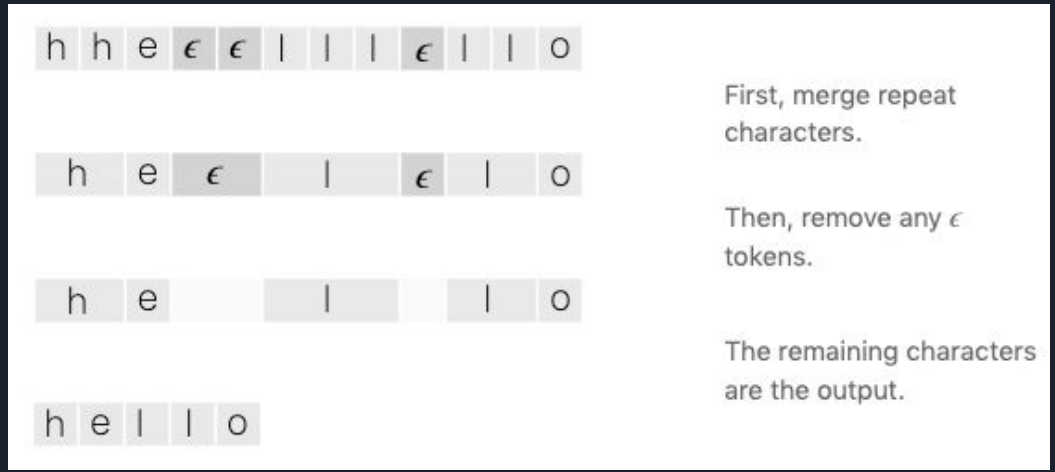

- 6300 audio samples
- 630 speakers
- ~10 hours
- 2 Genders, 8 Dialects

# Speech to Text: Model

```
Net(
  (lstm): LSTM(513, 128, num_layers=2, batch_first=True)
  (linear): Linear(in_features=256, out_features=10, bias=True)
)
```

- Predicts: text given speech signal
- Optimizer: Adam with weight decay
- Loss: Connectionist Temporal Classification (CTC)
- Gradient norm clipped at 400

# Speech to Text: CTC algorithm




Source: https://distill.pub/2017/ctc/

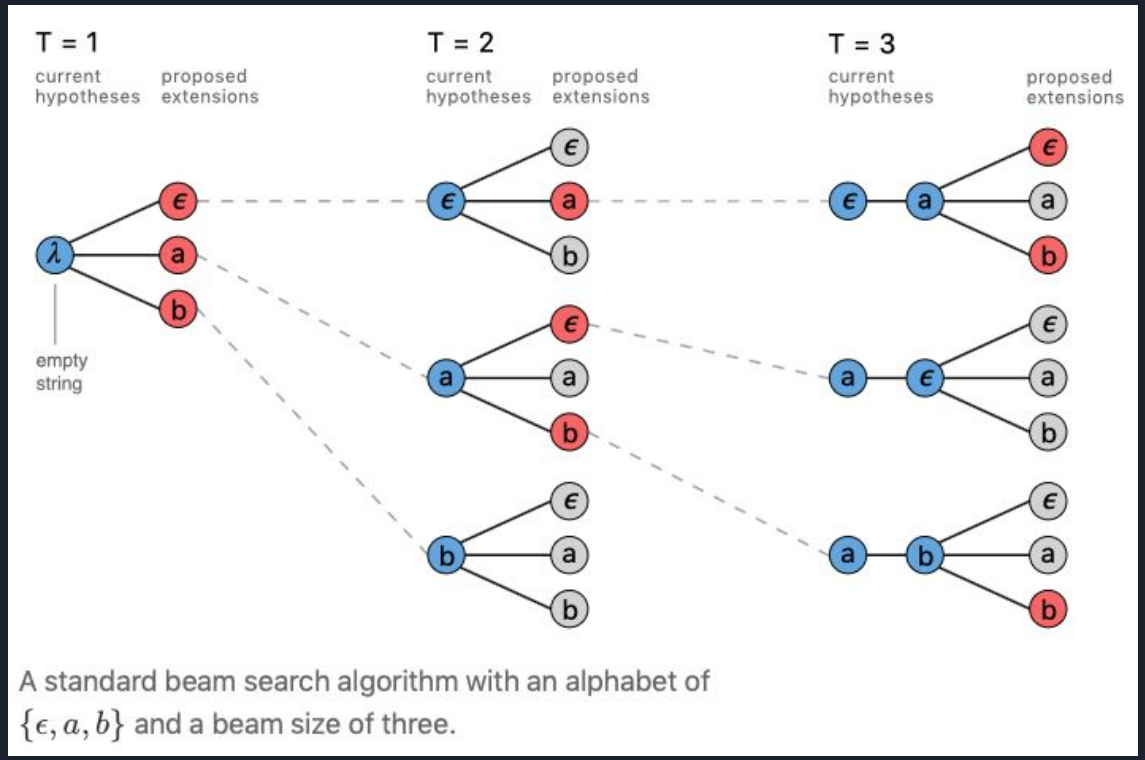
# Speech to Text: CTC loss

$$p(Y \mid X) = \sum_{A \in \mathcal{A}_{X,Y}} \prod_{t=1}^{T} p_t(a_t \mid X)$$

The CTC conditional **probability** **marginalizes** over the set of valid alignments computing the **probability** for a single alignment step-by-step.

Source: https://distill.pub/2017/ctc/

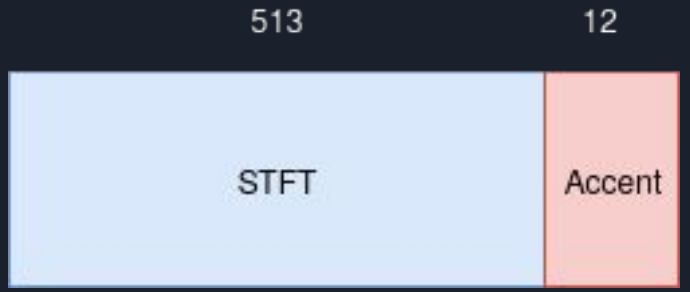
# Speech to Text: Beam search decoding




A standard beam search algorithm with an alphabet of $\{\epsilon, a, b\}$ and a beam size of three.

Source: https://distill.pub/2017/ctc/

# Speech to Text: Conditioning



# Speech to Text: Results

**Ground Truth:** *destroy every file related to my audits*
**Prediction:** *es proy evreyfo rtlathemotes*

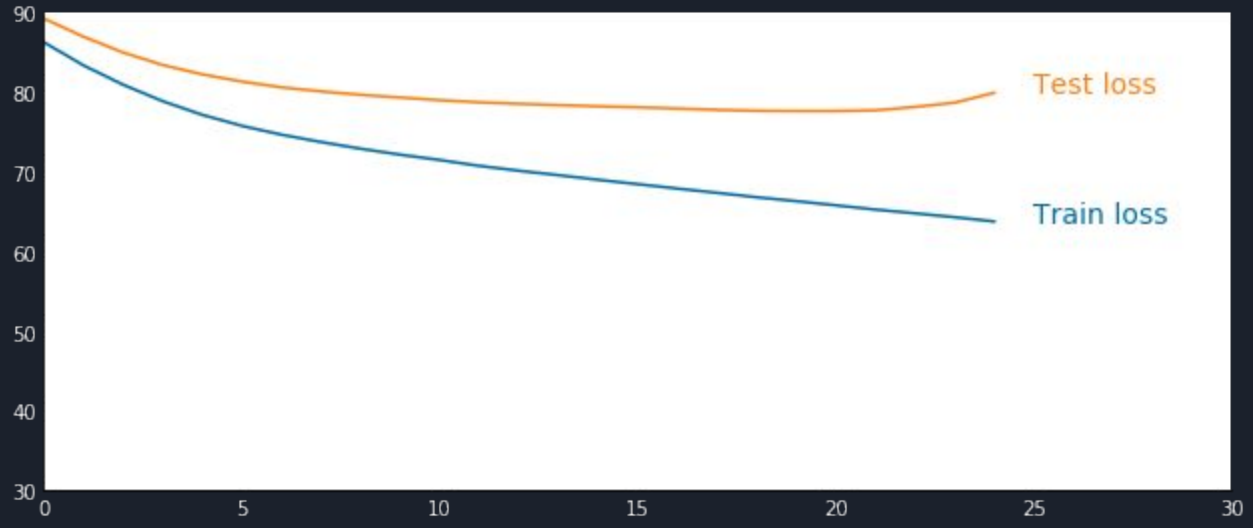
**Ground Truth:** *nobody in his right mind punishes a quarter century old dereliction*
**Prediction:** *non melit ias righ moting ciishous if alervershantre dar lit h*

**Ground Truth:** *the cranberry bog gets very pretty in autumn*
**Prediction:** *he crn brid bo dis puri thr e a or*

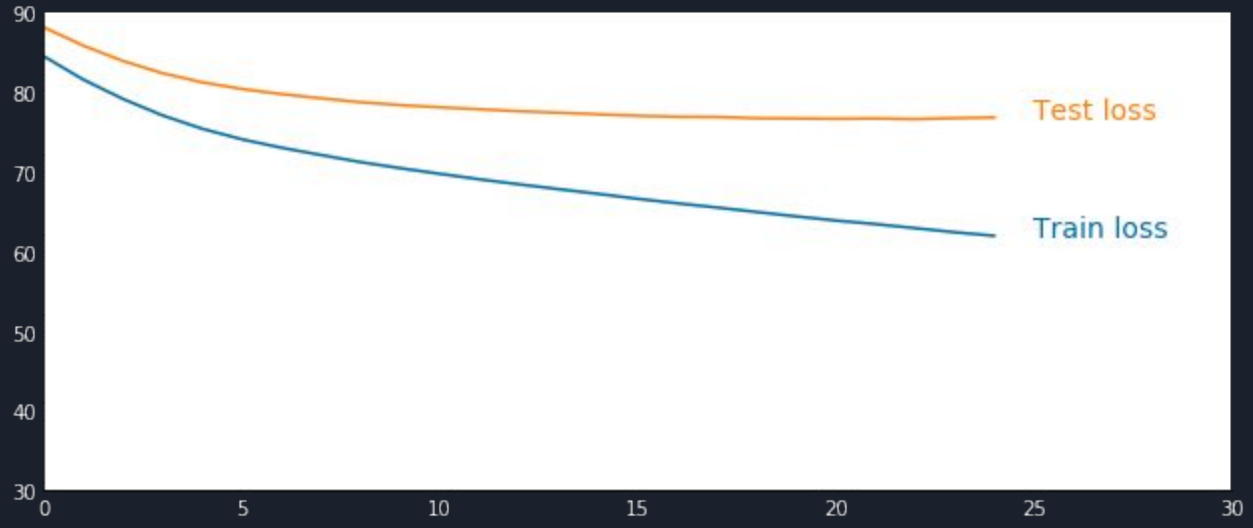
**Ground Truth:** *don't ask me to carry an oily rag like that*
**Prediction:** *don't ask me to cacarry an oily rag like that*

**Ground Truth:** *she had your dark suit in greasy wash water all year*
**Prediction:** *she had your dark suit in greasy wash water all year*

# Speech to Text: Learning Curve



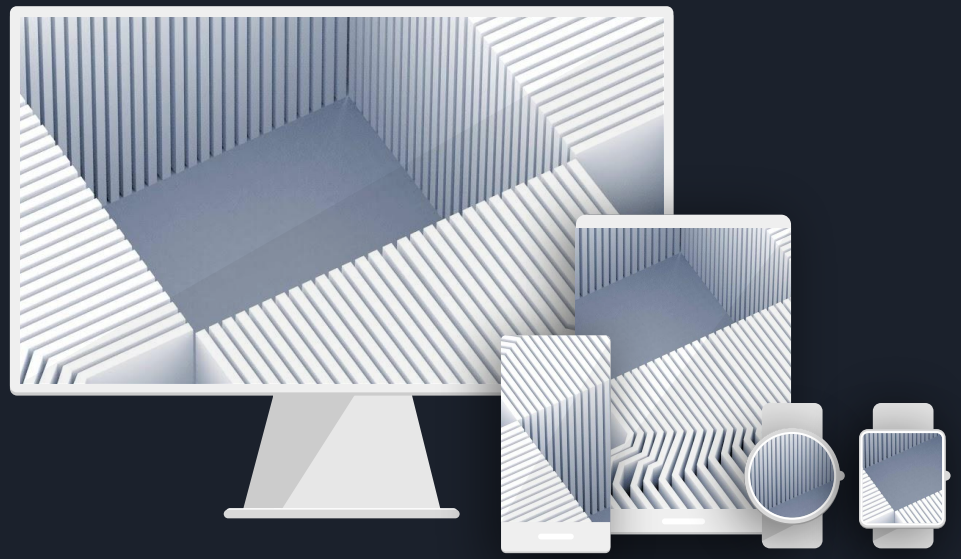
# Speech to Text: Learning Curve with accent

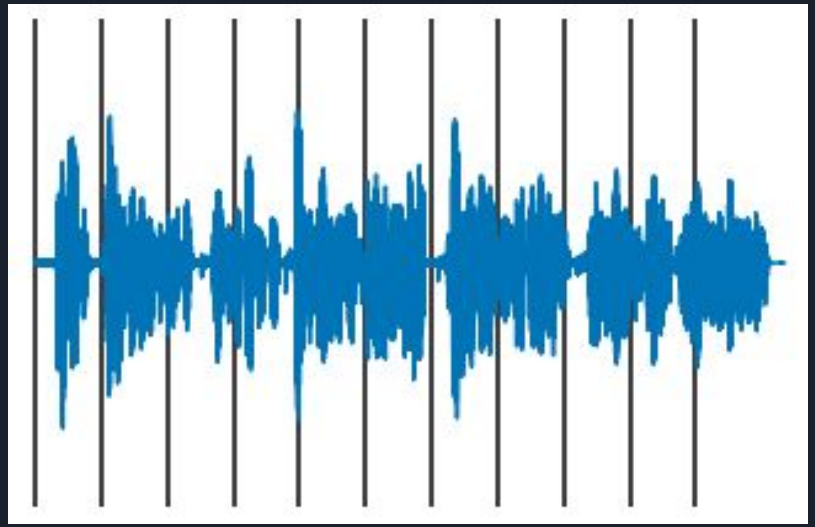


# Next steps

- Run the NN on Mozilla Common Voice dataset
- Experiment with other network architectures, such as:
  - Siamese network for classification
  - 1-D CNN for classification and sequence prediction
  - Beam search for decoding
  - Conditional concatenation using output embeddings from the classification model
- Try out other audio transformations, such as MFCC

Thank you!





Splitting long paragraphs into smaller chunks