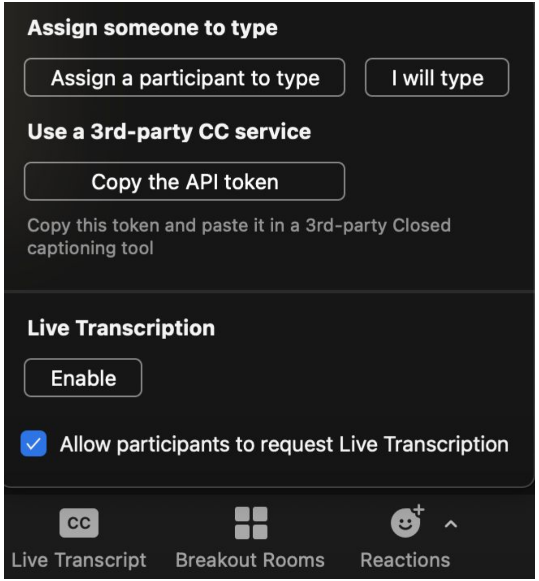
Wav442Letter:

an efficient speech-to-text system

Matt Palazzolo, Andrew Schallwig, Aditya Singhvi

Motivation



Assign someone to type

Assign a participant to type | I will type

Use a 3rd-party CC service

Copy the API token

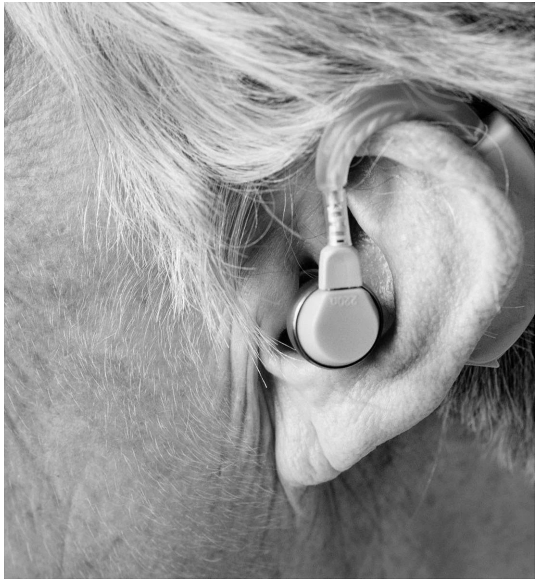
Copy this token and paste it in a 3rd-party Closed captioning tool

Live Transcription

Enable

☑ Allow participants to request Live Transcription

Auto-Captioning

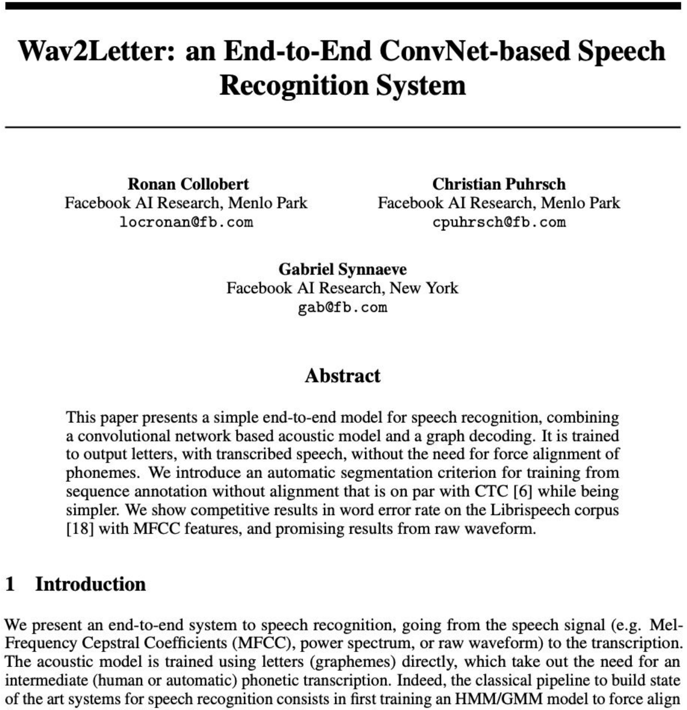


Accessibility

Voice-Controlled Devices

Related Work

- Fully-convolutional network
- Audio File → Transcript
- Introduced new, faster criterion for decoding classification to text transcript
- State-of-the-art results:
  - Word-error-rate (WER): 7.2%
  - Letter-error-rate (LER): 6.9%
- Can we recreate the results with far less data and far less computational power?



Wav2Letter: an End-to-End ConvNet-based Speech Recognition System

Ronan Collobert
Facebook AI Research, Menlo Park
locronan@fb.com

Christian Puhrsch
Facebook AI Research, Menlo Park
cpuhrsch@fb.com

Gabriel Synnaeve
Facebook AI Research, New York
gab@fb.com

Abstract

This paper presents a simple end-to-end model for speech recognition, combining a convolutional network based acoustic model and a graph decoding. It is trained to output letters, with transcribed speech, without the need for force alignment of phonemes. We introduce an automatic segmentation criterion for training from sequence annotation without alignment that is on par with CTC [6] while being simpler. We show competitive results in word error rate on the Librispeech corpus [18] with MFCC features, and promising results from raw waveform.

1 Introduction

We present an end-to-end system to speech recognition, going from the speech signal (e.g. Mel-Frequency Cepstral Coefficients (MFCC), power spectrum, or raw waveform) to the transcription. The acoustic model is trained using letters (graphemes) directly, which take out the need for an intermediate (human or automatic) phonetic transcription. Indeed, the classical pipeline to build state of the art systems for speech recognition consists in first training an HMM/GMM model to force align

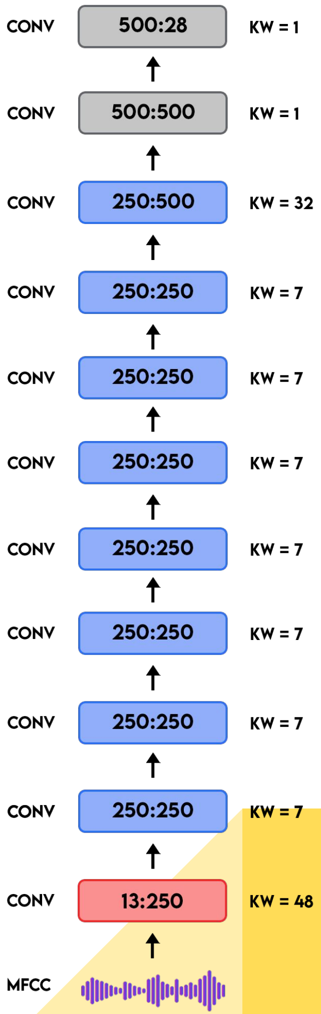
Wav2Letter (Collobert & Puhrsch, 2016)

Librispeech Dataset (OpenSLR)

- 1000 hours of audiobook snippets with transcription
- We use 2703 samples in *dev-clean* portion (~5.4 hours)
    - sort data → split into uniform batches → pad within batches
- Audio Encoding: Mel-Frequency Cepstral Coefficients (MFCC)
    - 2D Tensor of *[MFCC features, Time]*
- Text Encoding: 1D Tensor
    - 28 classes: English alphabet, apostrophe (‘), space
- https://www.openslr.org/12

Wav442Letter Architecture

- Last two layers act as fully-connected network (Kernel Width = 1)
- Modifications from Wav2Letter:
    - ○ 2000 → 500 channels in layers 9 - 11 (75% reduction in trainable parameters)
    - ○ Investigated more convolutional layers (2 - 8) with smaller kernels



Decoding Classifications (CTC)

- Problem: Each "frame" of the input audio (~12.5 ms) will have a single class prediction (i.e. letter) — how do we transform this into a transcript to calculate loss with the ground truth?
- Solution: CTC Loss (Graves et al. 2006)

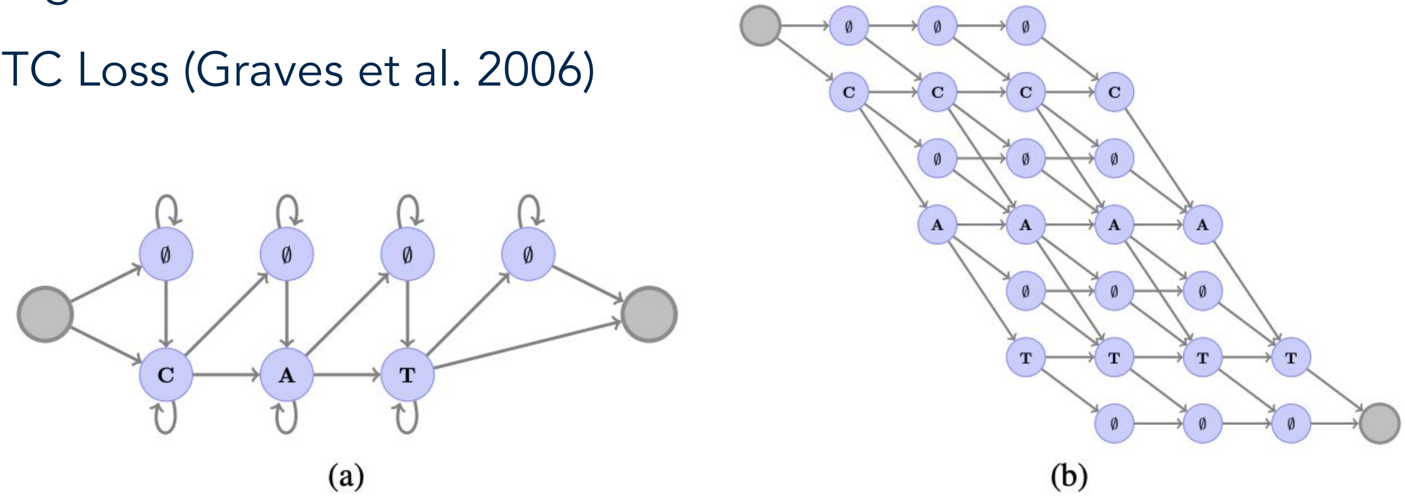


Figure 2: The CTC criterion graph. (a) Graph which represents all the acceptable sequences of letters (with the blank state denoted "∅"), for the transcription "cat". (b) Shows the same graph unfolded over 5 frames. There are no transitions scores. At each time step, nodes are assigned a conditional probability output by the neural network acoustic model.

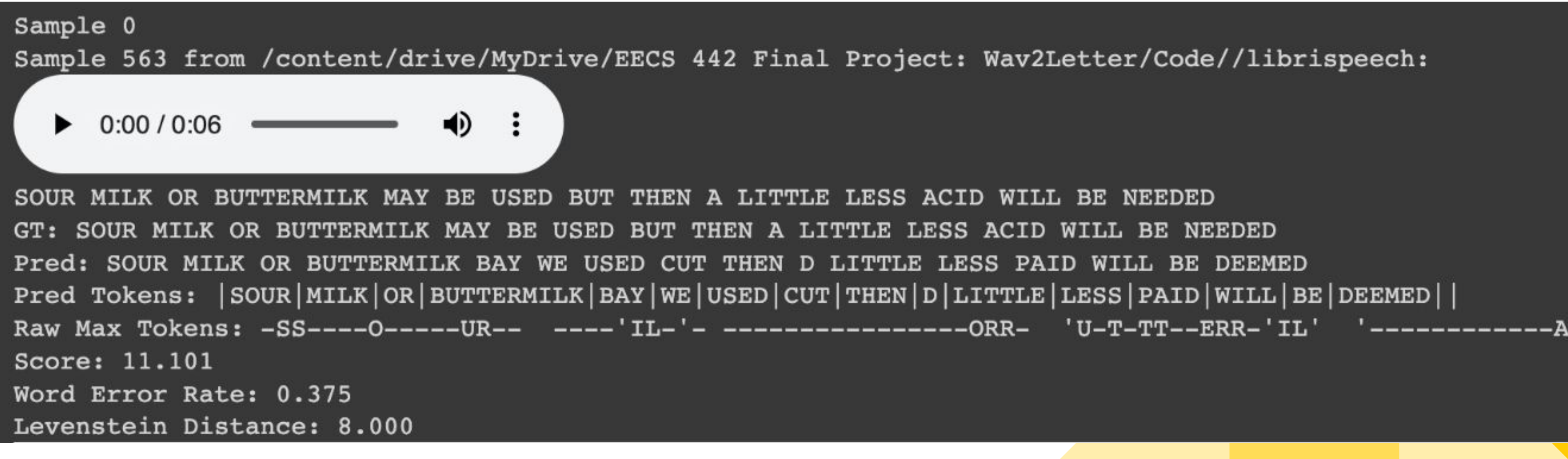
Results

	Wav2Letter	Wav442Letter
Train Dataset Size	960 hours	3.7 hours
Learnable Parameters	23,282,529	7,486,029
Word-Error-Rate (WER)	7.2%	40.1%
Letter-Error-Rate (LER)	6.9%	19.1%

Results

Sour milk or buttermilk may be used but then a little less acid will be needed

Sour milk or buttermilk bay be used cut then d little less paid will be deemed

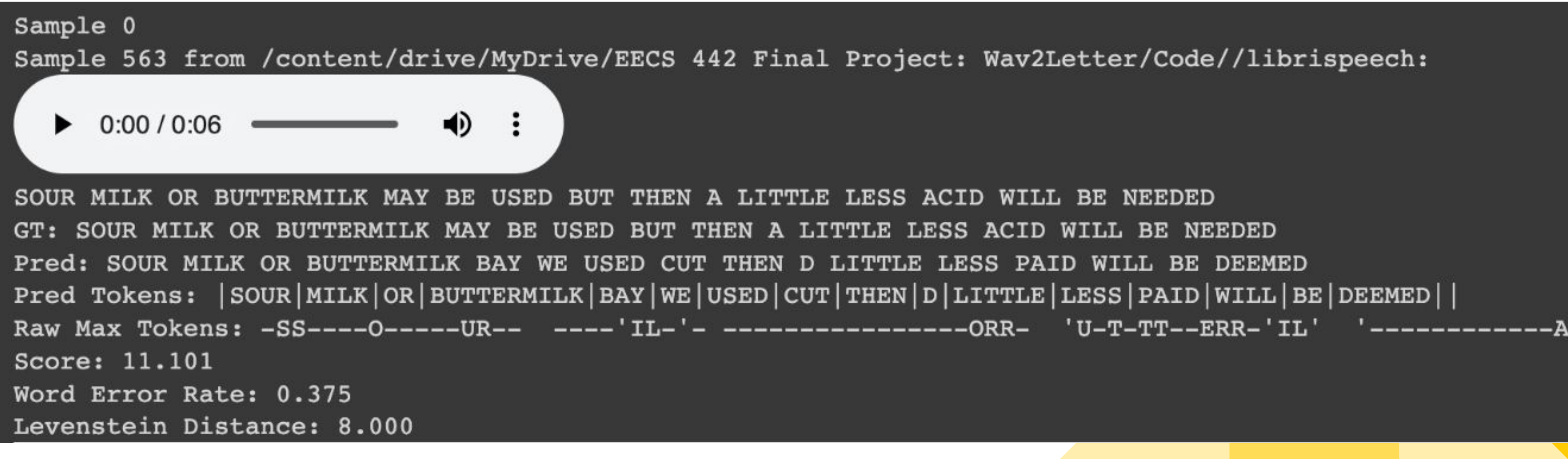


```
Sample 0
Sample 563 from /content/drive/MyDrive/EECS 442 Final Project: Wav2Letter/Code//librispeech:
```

```
SOUR MILK OR BUTTERMILK MAY BE USED BUT THEN A LITTLE LESS ACID WILL BE NEEDED
GT: SOUR MILK OR BUTTERMILK MAY BE USED BUT THEN A LITTLE LESS ACID WILL BE NEEDED
Pred: SOUR MILK OR BUTTERMILK BAY WE USED CUT THEN D LITTLE LESS PAID WILL BE DEEMED
Pred Tokens: |SOUR|MILK|OR|BUTTERMILK|BAY|WE|USED|CUT|THEN|D|LITTLE|LESS|PAID|WILL|BE|DEEMED||
Raw Max Tokens: -SS----O-----UR--  ----'IL-'- ----------------ORR-  'U-T-TT--ERR-'IL'  '-----------A
Score: 11.101
Word Error Rate: 0.375
Levenstein Distance: 8.000
```

Results

Sour milk or buttermilk bay be used cut
then d little less paid will be deemed



```
Sample 0
Sample 563 from /content/drive/MyDrive/EECS 442 Final Project: Wav2Letter/Code//librispeech:
```

```
SOUR MILK OR BUTTERMILK MAY BE USED BUT THEN A LITTLE LESS ACID WILL BE NEEDED
GT: SOUR MILK OR BUTTERMILK MAY BE USED BUT THEN A LITTLE LESS ACID WILL BE NEEDED
Pred: SOUR MILK OR BUTTERMILK BAY WE USED CUT THEN D LITTLE LESS PAID WILL BE DEEMED
Pred Tokens: |SOUR|MILK|OR|BUTTERMILK|BAY|WE|USED|CUT|THEN|D|LITTLE|LESS|PAID|WILL|BE|DEEMED||
Raw Max Tokens: -SS----O-----UR--  ----'IL-'- ----------------ORR-  'U-T-TT--ERR-'IL'  '-----------A
Score: 11.101
Word Error Rate: 0.375
Levenstein Distance: 8.000
```

Questions