CS230

Speech to Text System: Pastor Wang Mandarin Bible Teachings (Speech Recognition)

Yen Peng (Karl) Kao*
Department of Computer Science
Stanford University
karlkao@stanford.edu

Abstract

A Bible study group needs a speech-to-text system to transcribe Pastor Wang's teachings in Chinese mandarin. We trained an end-to-end deep learning model, and applied strategy to lower error rate of speech recognition from 50.39% to 24.97%.

1 Introduction

Motivation of this system comes from an aspiration to help a large study group ease everyday life. The group aims to read through The Holy Bible at pace of one chapter a day along with a pastor's teachings. Pastor Wang has been recording audio, chapter-by-chapter teachings based on the Bible of Chinese Union Revision [1] . Fellow of this group is going on a long journey of 1189 chapters of Bible reading and hundreds of hours of Bible teachings in Chinese Mandarin.

Problem the group faces is content of the audio teachings is not searchable. The group fellow feels helpless and not able to find information need in an efficient manner. Automatic Speech Recognition (ASR) technique may assist in a way to transcribe audio content into searchable text without human intervention. The technique is therefore being applied to the system to document such abundant Bible teaching materials.

Target of our system is unique. We focus on biblical context and current ASR model doesn't show relevant results about any pastor's Bible teachings in mandarin. This study aims to investigate practical use case which limits scope to a single speaker, Pastor Wang. Existent ASR system may not generalize well for utterance in Bible context. The ASR likely misses Bible terminology and coherence of Bible teachings. The terminology is for example birthplace of Jesus, Bethlehem. Teaching coherence indicates main topic of a talk stays consistent throughout the audio recording, such as teaching about birth of Jesus. This study is initiated to investigate the generalization gap between current model and Chinese Bible teachings recorded by Pastor Wang. The goal is to understand the gap and come up with practical methodology to mitigate the gap.

Input to our ASR system is an audio clip of Pastor Wang's Bible teachings in Chinese mandarin. We then run an end-to-end deep learning model to output transcript of the audio clip. Character Error Rate (CER) is a single metric to measure accuracy of the transcribing task. We found data is a major

*SCPD student, NVIDIA full-time employee, can be reached at email: kakao@nvidia.com.

CS230: Deep Learning, Winter 2020, Stanford University, CA.

factor to achieve satisfactorily low CER. Computation resources along with hyperparameters tuning process are key to efficiently deliver an effective ASR system.

2 Related work

Attributed to quick development of deep learning technology, some of ASR models are productized [2][3][4][5] in market, and claiming with high accuracy of speech-to-text recognition outcome. These products support the speech-to-text feature in Chinese mandarin and may be used to solve the problem our system is trying to address.

The AWS system [2] achieves state-of-the-art accuracy with CER of 3.83% using our test case. However, the system doesn't disclose much technical details. We therefore use 3.83% CER as benchmark and take open source [6] repository, an implementation of DeepSppech2 model [3], as a development platform. The work [7] further provides insight into mandarin around the DeepSpeech2 model.

3 Dataset and Features

This study uses dataset (aishell) from OpenSLR [8] as first training set. The aishell is corpus of mandarin Chinese encoded in audio wav format (16kHz, 16-bit). We later on got academic license to use a larger dataset (aishell2) [9] in the same audio format as aishell. Pastor Wang's Bible Teachings is dataset we aim to have audio recording transcribed into text. In Table 1, we list features of these dataset and size of the labeled data. Wang's Teachings are labeled by our own and because of limited amount of the labeled data it's being used as Test Set to measure CER metric.

Consistency of audio coding is crucial for ASR system to perform. We do data preprocessing and regulate Test Set to fit into our ASR model. The Test Set are reformatted from MP3 to wav (16kHz, 16-bit) matched coding of Training Set. FFmpeg [10] works like a Swiss knife and it is able to convert various audio formats. Audio recordings are further split into multiple clips with recording length between 1 to 27 seconds. The audio segmentation shall fit into model constraint. We use Voice Activity Detection (VAD) [11] to split long recordings.

The Figure1 illustrates a clip of audio waveform which encodes eight Chinese mandarin characters. Encoded Transcription: [1]讲[2]到[3]神[4]伟[5]大[6]的[7]主[8]权

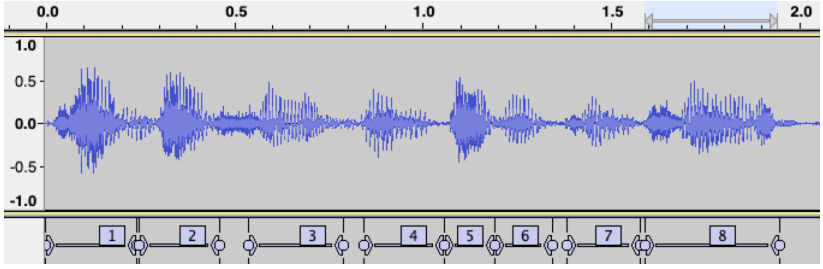

Figure 1: Audio waveform encoding eight Chinese mandarin characters, approximately 2 seconds long

4 Methods

This study takes DeepSpeech2 model [3] and its implementation [6] as development platform. This model is trained with the Connectionist Temporal Classification (CTC) loss function [12] to predict speech transcriptions from audio. Figure 2 illustrates model architecture. The architecture may

	Aishell		Pastor Wang
	(aishell)	(aishell2)	(Teachings)
Dataset Size (utterance)	141,925	1,0009,222	226
Dataset Size (hour)	151	1,001	0.3
Training Set (%)	84 .8	99.0	0
Training-dev Set (%)	10.1	0.5	0
Training-test Set (%)	5.1	0.5	0
Test Set (%)	0	0	100

Table 1: Dataset Size and Division

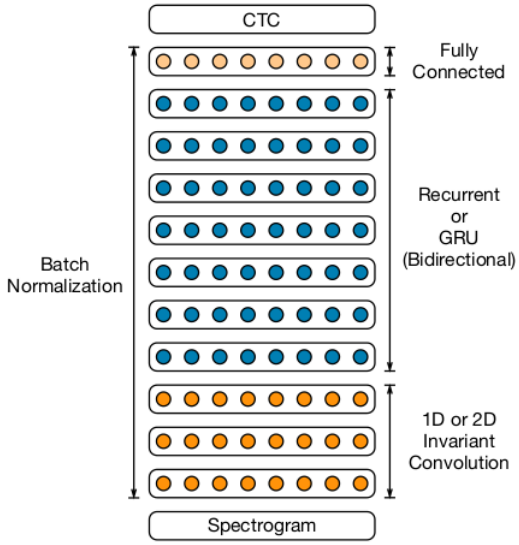

Figure 2: Deep Speech 2 Architecture, variants of this architecture include the number of convolutional layers from 1 to 3 and the number recurrent or GRU layers from 1 to 7

include the number of convolutional (Conv) layers from 1 to 3 and the nubmer of recurrent or GRU layers from 1 to 7. We explored 2-Conv, 3-GRU layers and compared it against the maximum capacity of the architecture 3-Conv, 7-GRU.

5 Experiments/Results/Discussion

The system initially trained Deepspeech2 with a labeled dataset (aishell), and also came across a pretrained Deepspeech2 model (baiducn1.2k). They both performed poorly (CER 50.39% and 54.30% respectively) in our test case. We then explored AWS Transcribe as a test vehicle and the system delivered nearly state-of-the-art CER (3.83%) with our Test Set. Human-level performance can achieve perfect CER (0.0%) against the Test Set in setup of single person validation. The AWS Transcribe satisfies us as a benchmark. Table 2 lists the DeepSpeech2 with two dataset and the benchmark from AWS Transcribe.

CER of DeepSpeech2 and benchmark respectively is 50.39% vs. 3.83%. The significant delta strongly indicates we can improve the DeepSpeech2 to achieve higher accuracy rate. We then performed error analysis to understand which area is to improve.

Error Analysis The Error Analysis performed on DeepSpeech2 model (aishell) indicates the trained model is not able to perform accurate character by character transcription in some cases.

	DeepSpeech2 (aishell)	DeepSpeech2 (baiducn1.2k)	AWS Transcribe (Unknown)
Language Model	zhidao_giga	zhidao_giga	Unknown
Dataset Available	Yes	No	No
Dataset Size (utterance)	141,925	Unknown	Unknown
Dataset Size (hour)	151	1,204	Unknown
Wang Test Set CER (%)	50.39	54.30	3.83

Table 2: DeepSpeech2 and AWS Transcribe Benchmark, Model with Training Dataset Size vs CER Results against Wang's Teachings Test Set

	DeepSpeech2			
Dataset	**(aishell2)**		**(aishell)**	
Dataset Size (utterance)	999,077		141,925	
Dataset Size (hour)	990		151	
Conv Layers	2	3	2	2
RNN Layers (GRU)	3	7	3	3
Training Epoch	50	50	50	50
Training Time (hour)	103.3	107.8	15.1	15 .1
Batch Size	16	16	16	16
Training Loss	0.188	0.069	0.009	0.007
Training-Dev Loss	5.284	5.011	10.900	13.112
alpha tuned best	2.6	2.2	4.2	2.6
bata tuned best	5.3	4.4	10.0	5.0
Wang Testing CER (%)	28.98	28.08	42.87	50.39

Table 3: DeepSpeech2 of Model Layout and alpha / beta Tuning vs Test with Wang's Teachings

The following example illustrates this failure mode with CER of 65.38%. The output length is even not matched with the target transcription length.
Target Transcription: 各位亲爱的弟兄姐妹大家好从今天呃这一讲开始我们会连着
Output Transcription: 昨日价格对行者的叫好从今天的这将开始我会原则
Character Error Rate [CER]: 65.38%

Another example illustrates the trained model perform poorly on recognizing Pastor Wang's utterance and it might be attributed to his vocal accent with CER of 62.50%. The output might be sounded similar to recording of the target transcription.
Target Transcription: 讲到神伟大的主权
Output Transcription: 将到人轨道的主持
Character Error Rate [CER]: 62.50%

CER Improvement Based on the conducted error analysis, we found in such end-to-end deep learning model the used training dataset [8] might not be large enough to achieve benchmark CER. First thing trying to improve CER is to get a larger dataset and we fortunately license a larger dataset, ashiell2 [9]. We also tune hyperparameter alpha and beta. By creating an alpha / beta grid, we find the best set of these two variables, though it takes time to iterate selected combinations. With aid of the larger dataset and hyperparameter tuning, the system dropped more than 20% CER. Last attempt, we tried to build a larger neural network to further improve CER. However, this approach does not give a promising result. Table 3 shows all the attempts and results. Table 4 shows respective CER with respect to training epoch.

The improved model almost matches character count with more accurate transcription, CER down from 65.38%, compared to the previous model with smaller training set.
Target Transcription: 各位亲爱的弟兄姐妹大家好从今天呃这一讲开始我们会连着
Output Transcription: 更是前一个地区准备大家好从今天的这讲开始我们会连折
Character Error Rate [CER]: 46.15%

DeepSpeech2 (aishell2)				
Conv Layers	2	2	2	2
RNN Layers (GRU)	3	3	3	3
Training Epoch	10	20	30	50
Training Time (hour)	19	41.3	58.3	103.4
Batch Size	16	16	16	16
Training Loss	1.601	0.612	0.065	0.188
Training-Dev Loss	4.057	4.454	5.498	5.284
alpha tuned best	2.1	2.6	3.3	2.6
bata tuned best	4.2	4.7	4.4	5.3
Wang Testing CER (%)	24.97	25.73	26.09	28.98

Table 4: Training epoch effects with best-tuned set of alpha and beta vs Test with Wang's Teachings

The improved model transcribes speech better, CER down from 62.50%, compared to the previous model with smaller training set.
Target Transcription: 讲到神伟大的主权
Output Transcription: 讲到神鬼道的主权
Character Error Rate [CER]: 25%

Training Efficiency Improvement Attempt We attempted to accelerate training per epoch by using multiple GPUs, blessed with NVIDIA DGX-2 server [13]. Unfortunately, this model does not accelerate while using multiple GPUs as Figure 3 shows.

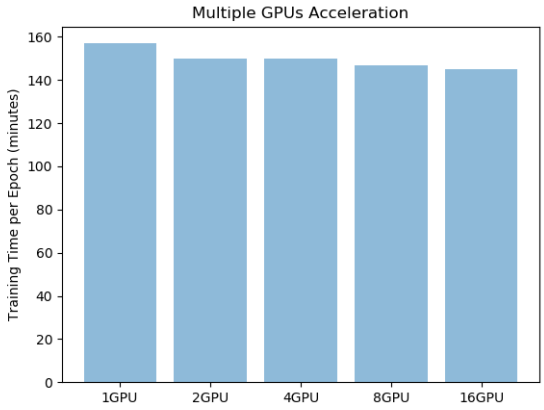

Figure 3: Attempt Multiple GPUs to Accelerate Epoch Training

Once attempting to increasing mini-batch size larger than 16 or increasing the neuron number of RNN from 1024 to 2048, we ran into out of memory error. For example,

```
Out of memory error on GPU 7. Cannot allocate 81.797119MB memory on GPU 7,
available memory is only 29.437500MB.
```

6 Conclusion/Future Work

Applying deep learning technique to develop ASR system is an elaborate engineering task. Large training dataset, data preprocessing, fine tuning hyperparameter, proper model architecture, avoiding training overfit, and even data log management are all contributing to a successful system in our study. We may continue to identify data mismatch to lower CER in our application and address attempts on training efficient for future work.

References

[1] Calvin Mateer et al. The bible, chinese union version. China Christian Council or Hong Kong Bible Society, 1919.

[2] Amazon. Aws transcribe.

[3] Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, Jie Chen, Jingdong Chen, Zhijie Chen, Mike Chrzanowski, Adam Coates, Greg Diamos, Ke Ding, Niandong Du, Erich Elsen, Jesse Engel, Weiwei Fang, Linxi Fan, Christopher Fougner, Liang Gao, Caixia Gong, Awni Hannun, Tony Han, Lappi Johannes, Bing Jiang, Cai Ju, Billy Jun, Patrick LeGresley, Libby Lin, Junjie Liu, Yang Liu, Weigao Li, Xiangang Li, Dongpeng Ma, Sharan Narang, Andrew Ng, Sherjil Ozair, Yiping Peng, Ryan Prenger, Sheng Qian, Zongfeng Quan, Jonathan Raiman, Vinay Rao, Sanjeev Satheesh, David Seetapun, Shubho Sengupta, Kavya Srinet, Anuroop Sriram, Haiyuan Tang, Liliang Tang, Chong Wang, Jidong Wang, Kaifu Wang, Yi Wang, Zhijian Wang, Zhiqian Wang, Shuang Wu, Likai Wei, Bo Xiao, Wen Xie, Yan Xie, Dani Yogatama, Bin Yuan, Jun Zhan, and Zhenyao Zhu. Deep speech 2 : End-to-end speech recognition in english and mandarin. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 173–182, New York, New York, USA, 20–22 Jun 2016. PMLR.

[4] Google. Gcp speech-to-text.

[5] Microsoft. Azure speech-to-text.

[6] https://github.com/PaddlePaddle/DeepSpeech. Deepspeech2 on paddlepaddle.

[7] Ryan J. Prenger and Tony Han. Around the world in 60 days: Getting deep speech to work in mandarin.

[8] Beijing Shell Shell Technology. Mandarin data (aishell).

[9] Beijing Shell Shell Technology. Mandarin data (aishell2).

[10] FFmpeg Org. Ffmpeg a collection of libraries and tools to process multimedia content.

[11] John Wiseman. Webrtc voice activity detector (vad).

[12] Alex Graves, Santiago Fernández, and Faustino Gomez. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In *In Proceedings of the International Conference on Machine Learning, ICML 2006*, pages 369–376, 2006.

[13] NVIDIA Corp. Nvidia dgx-2, the world's most powerful ai system for the most complext ai challenges.