A Novel Method for Dataset Creation for Deep Learning in Automatic Speech Recognition (ASR)

Andrew Smith
Department of Computer Science and Engineering
University of South Carolina
Columbia, SC 29208 USA
aks3@email.sc.edu

KEYWORDS

sleep-scoring, machine learning, artificial intelligence, neuroscience, electrophysiology

ACM Reference Format:
Andrew Smith. 2022. A Novel Method for Dataset Creation for Deep Learning in Automatic Speech Recognition (ASR). In *Proceedings of (BCB '22).* ACM, New York, NY, USA, 3 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION AND PROBLEM

Human-Computer Interaction (HCI) is a critical, highly-studied discipline concerned with the "design, evaluation and implementation of interactive computer systems for human use" [7]. With the emergence of the internet came the rise in the use of computers for humans to communicate with one another. The main method for humans to input text into a computer has always been the keyboard. The keyboard has the following limitations: low bandwidth, fixed geospatial location (requiring a user to sit or stand in one spot during use), and utilization of both of a user's hands (as opposed to the hands being free during use). Therefore, the desire for a new method of input is evident. Automatic Speech Recognition (ASR) is the use of a computer to automatically transcribe speech recorded with a microphone into digitized text. Both academia and the private sector have devoted a large portion of resources to the ASR problem and with great success. There are many ASR technologies available that vary across dimensions of complexity, quality, and cost. Most smartphone users have state-of-the-art (SOTA) ASR technology in their pockets. Improved ASR would change the lives of everyone who uses their smartphone for text messaging, emailing, or tweeting. Not only would ASR change smartphone users' lives, but everyone who visits restaurants, businesses, hotels, or any institution where interaction with an intelligent agent could be replaced by an intelligent system capable of ASR. The creation of labeled datasets for supervised learning in ASR is time-consuming and subject to variability. Therefore, a new methodology for automatically labeling speech data is necessary.

BCB '22, August 07–10, 2012, Chicago, IL

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/XXXXXXX.XXXXXXX

2 RELATED WORK

Deep Speech 2 is a set of speech recognition models due to Google [1]. Deep Speech 2 is an End-to-End automatic speech recognition system without hand-engineered features. One dataset used to train Deep Speech 2, for example, is the Wall Street Journal (WSJ) dataset. The WSJ dataset consists of read articles from the Wall Street Journal newspaper. The creation of these datasets is time-consuming and costly due to their centralized labeling paradigm. The labeling of these datasets does not scale well with an increasing number of individuals uploading articles because the uploading individuals are not the ones labeling the articles.

Before Deep Speech 2, the current state of the art in automatic speech recognition, various feed-forward models were explored [2], and also recurrent networks with convolution [**article:3**].

Another issue that arises with automatically labeling audio input is having variable length sequences as well as variable length input. Therefore, one of the most critical aspects of this work is how to address that. Commonly, the CTC loss function [**article:2**] with a recurrent neural network is used to model distorted temporal information. Popularly Deep Speech 1 and 2 have used the CTC loss function to capture these temporal alignments.

In addition to model architecture and variable length sequences, in recent years, the amount of data has been critical for the success of state-of-the-art automatic speech recognition models. For example, Deep Speech 1 was trained on over 7000 hours of labeled speech [5].

3 APPROACH

Here we propose and evaluate a novel method for dataset creation for deep learning in the context of automatic speech recognition. In order to evaluate the novel dataset, we train a model with an architecture similar to Deep Speech 2.

YouTube is the most popular repository for video transactions between creators and consumers. Approximately 500 hours of video are uploaded to YouTube every minute [3].

Tools exist for downloading isolated audio from YouTube videos. If the creator of a video has uploaded a transcript with their video, YouTube provides the options for closed captions. Further, the transcript for any YouTube video with closed captions is downloadable. The format of the transcript comes with two features: timestamps and phrases. Each timestamp is paired with a phrase such that the phase occurs after each timestamp until the following timestamped phase begins. After many videos and corresponding transcripts are downloaded, they will be partitioned into phrases according to the associated transcript. This partitioning will produce what constitutes a labeled dataset ready for training a supervised machine

learning classifier. Discrete Fourier Transformation will be used to produce spectrograms for each phrase, clearly representing frequency content for the supervised machine learning classifier. An example input sample, titled with its corresponding target, paired with its corresponding input audio wave is shown in Fig 1. In total,

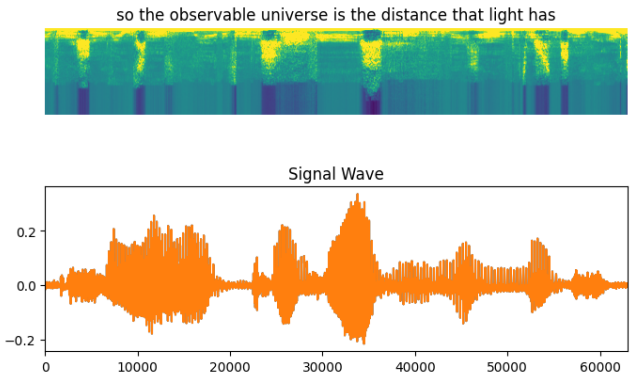

Figure 1: Sample input audio file paired with its spectrogram and titled by the target text sequence.

3 YouTube videos were taken for the creation of the data set for this investigation totaling 10 hours of data. The 3 YouTube videos were in podcast format, each from a different source. After partitioning the audio files that were extracted from the YouTube videos, there were approximately 12000 input-target pairs for training the model.

Importantly, the Connectionist Temporal Classification (CTC) algorithm [4] will be used to predict characters based on the unsegmented speech data. We have a dataset that has audio clips and transcripts. However, we do not know how the transcripts align with the audio clips. CTC has been shown to overcome this lack of information in automatic speech recognition. Therefore, understanding and implementing CTC will be pivotal to using such a dataset.

The architecture of the model used to test the novel data set from YouTube was modeled after that of Deep Speech 2, as shown in Fig. 2. First, a spectrogram is calculated, as shown in Fig 1, then a couple of convolution layers, a certain number of recurrent layers, a fully-connected layer, and finally the CTC loss function. In the architecture used for this work, there were approximately 27 million trainable parameters. With these many parameters, training time becomes highly non-trivial, even on our machine with two state-of-the-art consumer GPUs (NVIDIA 3090Ti).

3.1 Evaluation

The industry standard for evaluating a speech-to-text model is Word Error Rate (WER) which is a common, well-defined accuracy metric [6]. WER is derived from the Levenshtein distance, a classical distance based on the number of edits required to change one string into another using insertion, deletions, and substitutions. We will determine training and testing WER and CER for our model on our dataset and note how it compares to state-of-the-art word error rates and character error rates. Not only can we use testing data from YouTube that has been left out of training, but we can make

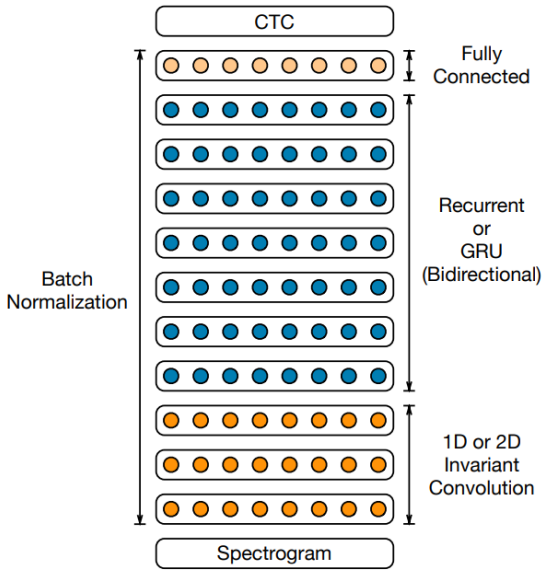

Figure 2: Architecture of Deep Speech 2 (cited previously), after which the architecture here was modeled to test the novel data set created from YouTube.

use of other labeled audio datasets such as movies with scripts or other labeled speech recognition datasets. Finally, we will use example target testing predictions as well as true values to convey how well the model is performing.

4 EVALUATION

After 50 epochs of training, the loss curve of the CTC loss function continues to decrease as shown in Fig 3. Deep Speech 2 has a

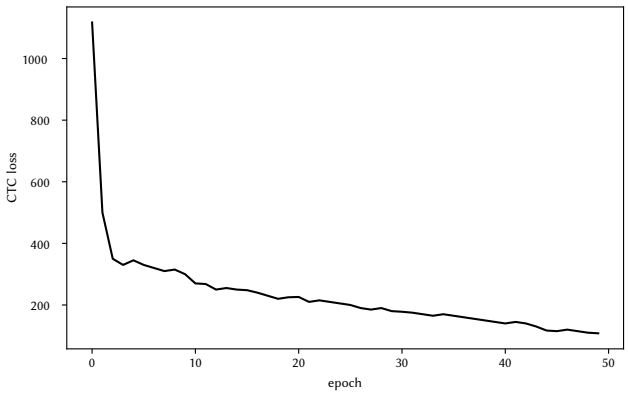

Figure 3: Training CTC loss over 50 epochs. Validation data was used in training but is not shown here.

comparably-sized test set, the "Baidu Test" which has approximately 3300 examples. Our test set used 10% of all examples, so it has 1200. Deep Speech 1 performed on this test set with a WER of 24.01, and Deep Speech 2 with a WER of 13.59. On our test set, which we

Test Set	DS1	DS2	Our Model
Baidu Test	24.01	13.59	
Our Test			66.9

Table 1: Performance of our model in terms of WER against baselines DS1 and DS2 on data sets of comparable size that are i.i.d.

would like to assume is independent and identically distributed to the "Baidu Test", our model performs with a WER of approximately 66.9. Therefore, against the baseline of Deep Speech 1 and Deep Speech 2 (also state-of-the-art), we perform, in terms of WER, as shown in Table 1. We will list now 5 target-prediction pairs to accentuate what information the model appears to be capturing from input audio sequences.

(1) Target: in eighteen eleven williams who murdered the marrs in ratcliffe highway having committed suicide in jail to escape hanging
(2) Prediction: in aite a leven wiloms wo murered the mars and relit hywa hing commisted suasid in jail to sca aniing
(3) Target: on his release an uncle a slopseller in chatham gave him a situation as barker or salesman
(4) Prediction: on his relys and oncal ay sloselr in shaterm ga himmasituation as barker or salsmon
(5) Target: even after sentence and until within a few hours of execution he was buoyed up with the hope of reprieve
(6) Prediction: the vin after senten sand intil within of forsobexcution he as boio with thu hal profrprve
(7) Target: oswald's marine training
(8) Prediction: oswalds moain traning
(9) Target: these materials are used by the animal in the manufacture of new protoplasm to take the place of that which has been used up
(10) Prediction: thes meterials ar us by the animol in the manyfactur of n protoplasm to tak the place of that which has beedusdu

5 DISCUSSION

5.1 Data set limitations

In comparison to the at least 7000 hours of audio used by Deep Speech 1 and Deep Speech 2, our 10 hours of audio is far behind. Having such a comparatively small data set likely hinders our model's ability to generalize speech patterns and filter out noise. The magnitude of our data set is not due to the novel data set creation technique, but due to the computational limitations imposed upon us. Instead of massive training units with teraflops of capability, we have one machine (though it has nice GPUs).

5.2 Computational Limitations

One training epoch here was observed to take approximately 1 hour. We trained our architecture for 50 epochs. This is equivalent to 50 hours of training time. Thus, the latency between training and testing imposes severe restrictions on experiment turnover time. Not only this, but the GPUs were each limited to approximately 8 gigabytes of memory; therefore, we had to use a small batch size to hold the model with 27 million parameters in memory. Should the size of the dataset increase, then time constraints would become intractable for this particular machine. Should the size of the model increase, then memory constraints would become intractable for this particular machine.

5.3 On the convergence of the model

At epoch 50, the model's loss had not quite converged. One could continue training the model and purportedly gain more performance. There might be a point where, for the given size of the data set, maximum performance is reached. However, it did not seem like that was reached here by observing the loss curve.

5.4 On novel data set creation

One objective of this work was to evaluate the efficacy of creating a data set using a novel method involving YouTube. However, it was not easy to tease out the effectiveness of this method because we trained a model here ourselves. If we had used a Deep Speech 2 version that is publicly available, we could have generated much more data and tested the model against that. Part of this constraint was due to the author's desire to learn how to use the CTC algorithm and understand the architecture of Deep Speech 2. After all, this work is a part of a course on the computer processing of natural language. Should this investigation develop into something more serious, the aforementioned error would be something for the next investigator to correct.

6 CONCLUSION

Here we evaluated a novel method for data set generation for training an automatic speech recognition model using end-to-end learning. The data set was generated using YouTube videos and their creator-uploaded transcripts. A model with an architecture similar to Deep Speech 2 was trained and evaluated on the novelly-created data set. It was delightful to see target-prediction pairs where the model seemed to be formulating something like text from audio files. The model clearly is extracting some true information from the input data.

REFERENCES

[1] Dario Amodei et al. 2015. Deep speech 2: end-to-end speech recognition in english and mandarin. (2015). DOI: 10.48550/ARXIV.1512.02595.
[2] H. Bourlard and N. Morgan. 1993. Connectionist speech recognition: a hybrid approach.
[3] Zippia Expert. 2022. How many hours of video are uploaded to youtube every minute?
[4] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd International Conference on Machine Learning* (ICML '06). Association for Computing Machinery, Pittsburgh, Pennsylvania, USA, 369–376. ISBN: 1595933832. DOI: 10.1145/1143844.1143891.
[5] Awni Hannun et al. 2014. Deep speech: scaling up end-to-end speech recognition. (2014). DOI: 10.48550/ARXIV.1412.5567.
[6] Jayashree Padmanabhan Melvin Jose Johnson Premkumar. 2015. Machine learning in automatic speech recognition: a survey.
[7] Hewett; Baecker; Card; Carey; Gasen; Mantei; Perlman; Strong; Verplank. 2014. "acm sigchi curricula for human-computer interaction". acm sigchi. (Aug. 2014).