האוניברסיטה העברית בירושלים
THE HEBREW UNIVERSITY OF JERUSALEM

Automatic Speech Recognition

Final Project For Introduction to Speech and Audio (67355)

Alon Ziv
Amit Shefi
Amit Roth (213212798)

Abstract

The goal of the project is to build the best ASR system we can using the tools we have learned in the course. There are many diverse ways to approach this problem: HMM, DTW, NN, etc. In this paper we will detail about the research and the experiments we have conducted towards developing the best ASR model we can.

Data

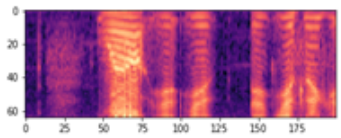
We will train the model on the AN4 dataset. This dataset is pretty small and consists of people spelling out addresses, names, letters, etc. We will utilize the size of the vocabulary in our decoding techniques.

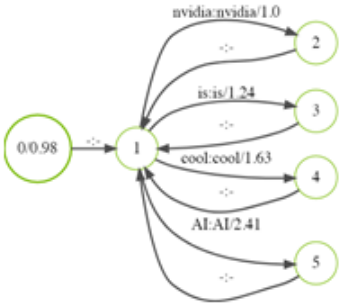
Our Approach

We decided developing an End-to-End system constructed from an acoustic model and a CTC decoder. We train the acoustic model on the data with CTC-loss and later infer the sentence using the decoder.

The acoustic gets the recording as a Mel spectrogram, an output for each time stamp a vector, which represents the probability for each token to appear. With this probability we infer the sentence using the decoder.

We started with a super naively one linear layer acoustic model and a greedy CTC decoder, and developed both acoustic model and decoder from this baseline. Our ASR system consists of two distinct components operating independently, prompting us to focus on enhancing each of these components individually. We developed additional 4 acoustic models and 2 CTC decoders. During the model evaluation, we have the flexibility to test each component in conjunction with the other and analyze the results. The research process proved fascinating and drove us into developing different aspects of the model - including the acoustic model, the decoder, and even augmenting data in order to train the acoustic model better, finally leading into out final model.





Acoustic Models

We will list all of the acoustic models we used in order, starting from the easiest model and infer it with each iteration.

1. linear layer:

   The simplest model. 1 linear layer with size $(n - mels, n - tokens)$

2. MiniLSTM
   - Input embedding - linear layer
   - 2 layer LSTM with hidden dimension 128
   - Readout - linear layerDee
3. DeepSpeech Toy - Based on the architecture of DeepSpeech Toy:
   - 3 Conv layers
   - 1 RNN layer
   - 2 Linear layers
4. DeepSpeech Small - Based on the architecture of DeepSpeech Small:
   - 2 Conv layers
   - 2 RNN layers
   - 2 Linear layers
5. DeepSpeech Large - Based on the architecture of DeepSpeech large:
   - 3 Conv layers
   - 4 RNN layers
   - 2 linear layers

Decoders

1. Greedy CTC - selects the most probable output label at each time step, without considering the context.

2. Beam Search Lexicon - This method explores multiple hypotheses simultaneously by considering a set of potential output labels at each time step and selecting the most likely paths based on a lexicon of all words available in the training data. The rationale of this model is that word that appeared in the data are more likely to reappear.
3. Beam Search LM - as the method mentioned above, but considering at the scoring mechanism also a language model (which have trained on different English data). This method has a large potential, but we will see that because of the differences in the data, the outcomes are not outstanding.

Training Process

Observing rapid overfitting caused by limited data, we introduced data augmentations to mitigate this issue and subsequently trained the models on the augmented dataset. Evaluating the models on both original and augmented datasets revealed significant and remarkable improvements.

DETAIL ABOUT THE AGUMENTATIONS

Benchmarks

In the following the table, we detail the WER values for all model combinations, data augmentations on different datasets:

Train, Test and Val

Acoustic Model / Decoder		Greedy	Lexicon	LM
Linear	NO	0.9878 0.9923 0.9916	0.9878 0.9930 0.9897	0.9842 0.9893 0.9832
	YES	0.9915 0.9940 0.9902	0.9905 0.9934 0.9863	0.9820 0.9919 0.984
LSTM	NO	0.7144 0.8168 0.8288	0.5814 0.6981 0.7195	0.6302 0.7365 0.7699
	YES	0.16777 0.4262 0.4885	0.0962 0.3252 0.3475	0.1115 0.3254 0.3472
DeepSpeech Tex	NO	0.0272 0.2911 0.3285	0.0043 0.2081 0.2054	0.0065 0.2 0.2132

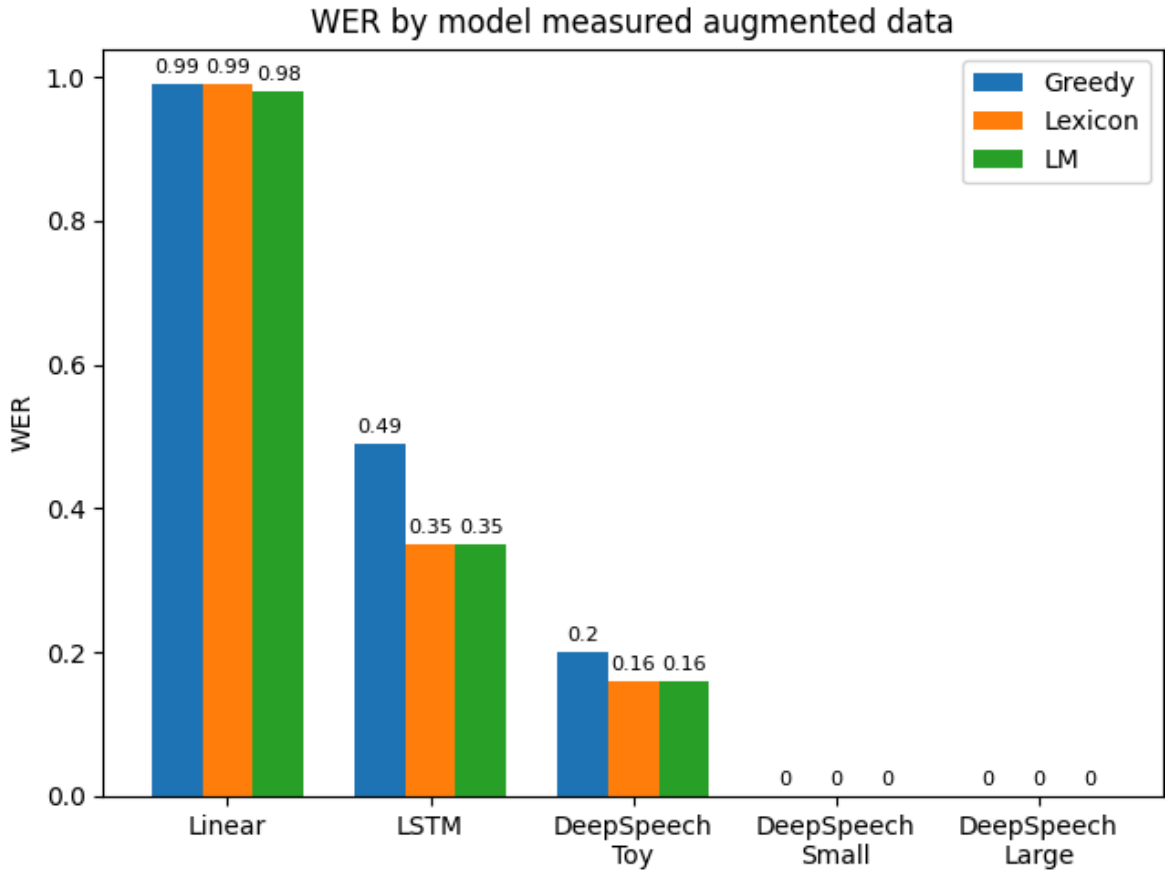
Acoustic Model / Decoder		Greedy	Lexicon	LM
DeepSpeech Toy	YES	0.0075 0.1869 0.2047	0.0086 0.1686 0.1625	0.0114 0.1862 0.1555
DeepSpeech Small				
DeepSpeech Big				

Graphs

Add graphs of the same data

Results

- Weak contribution of the language model for this data: not grammatically correct sentences, small vocabulary
- High contribution from augmentations



- BEST MODEL IS...

Future Work

?

Train our own language model maybe

Appendix

Git Repo - https://github.com/MajoRoth/ASR