Year 7
Final Project Report
2018-2019

7a.026-TAU - Improving Speech Recognition Robustness

A National Science Foundation
University Cooperative Research Center

Confidential and Proprietary to CVDI

7a.026-TUT - Improving Speech Recognition Robustness

Contents

Personnel 3
Executive Summary/Abstract 3
Goals and Objectives 4
Differences from Current State of Art 4
Methods and Datasets 5
Results 7
Functionality of Innovation(s) 8
Conclusions and Recommendations 8
Impact and Uses/Benefits 8
List of References 9
Appendix A 10

Confidential and Proprietary to CVDI

Personnel

PI First and Last Name (PI):

- Moncef Gabbouj

Other team member's first and last name (project role):

- Okko Räsänen (Co-PI)
- Ali Senhaji (Researcher)

Sponsoring IAB member's first and last name (company name):

- Silo.ai (Filip Ginter)

Executive Summary/Abstract

Speech recognition for general tasks is widely available in many languages by providers such as Google and Microsoft. However, current level technology has not been able to come up with a good quality general speech recognizer. Therefore, a good quality system requires training on case-specific datasets for the task at hand. This is not possible with e.g. Google Speech API. Another aspect of this customization is the ease with which it is possible to take additional aspects of the audio into account on subsequent models. Another concern with these cloud services is confidentiality. In some use cases, the data cannot be allowed to leave the organization in question. The costs of continued use of cloud services can also be considerable. This project aims to create an end-to-end solution for speech recognition. General deep learning-based speech recognizer is trained on open data and other available sources. The general model is used in the creation of better case-specific models, which are trained on client data. The resulting models are portable and can be set up in either cloud environments or local servers. The model can easily be combined with or serve as an input for additional ML models, such as sentence classification.

Robustness to noise and other interference in the environment is a crucial feature of a speech recognition system. The system should also be robust to different speakers, especially in public places, instead of being adapted specifically to each user. The purpose of this research project is to investigate different approaches to make speech recognition systems robust to noisy environments and different speakers. This project investigates speech enhancement techniques using the data-driven approaches.

Goals and Objectives

In this project, the goal is to implement a full Automatic Speech Recognition system. Our tasks were to:

- Identify the state-of-the-art architecture.
- Put together a new Dataset based on public domain resources for Automatic Speech Recognition task in Finnish language.
- Determine appropriate metrics for robustness.
- Benchmark different augmentation methods and propose a training strategy.

Differences from Current State of Art

We have used the DeepSpeech2 State of the art architecture with minor modifications. We have run few experiments and we have decided to use 2 Convolutional layers that take the speech spectrograms of the audio, then we have 5 hidden layers of type Gated Recurrent Unit (GRU) with a hidden size of 512 nodes and one fully connected layer at the end. The network is trained with the CTC activation function.

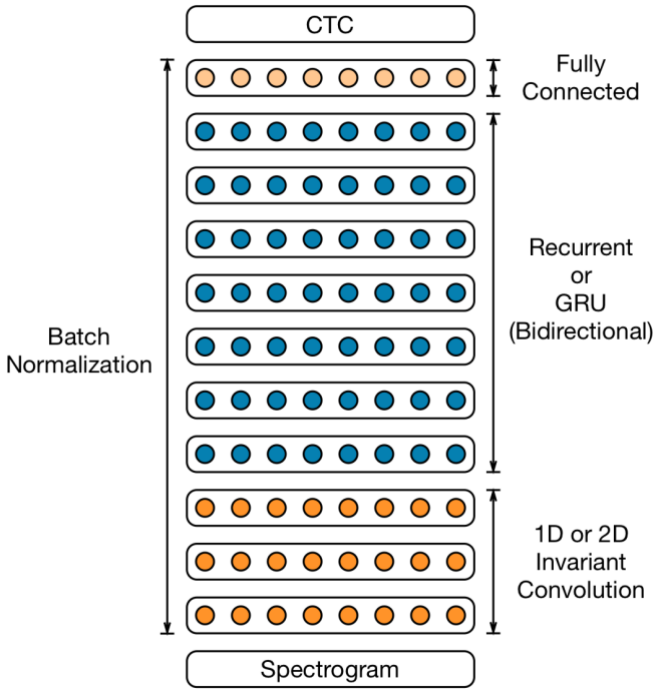

Figure 1. DeepSpeech2 Architecture

Confidential and Proprietary to CVDI

Methods and Datasets

Datasets

We used the LibriSpeech dataset. It is a corpus of approximately 1000 hours of 16kHz read English speech. The results of our experimentation are language independent. We are motivated to use this English dataset since it is widely used in the speech community and it was more practical to benchmark with the different baselines in the literature. We have limited ourselves to use just half of the training data to simulate learning a language with scarce data resources.

subset	hours	per-spk minutes	female spkrs	male spkrs	total spkrs
dev-clean	5.4	8	20	20	40
test-clean	5.4	8	20	20	40
dev-other	5.3	10	16	17	33
test-other	5.1	10	17	16	33
train-clean-100	100.6	25	125	126	251
train-clean-360	363.6	25	439	482	921
train-other-500	496.7	30	564	602	1166

Table 1. Statistics of the LibriSpeech Dataset

One of the main challenges faced when developing ASR systems for foreign languages is access to large corpus of annotated speech. We have also collected more than 50 hours of audiobooks that we put together from public domain in Finnish language.

Assessing Robustness

We assess robustness of a model by inferring different folds of the same testing data with different levels of SNR, starting with the clean test version and adding interferences and background noises with increased signal-to-interference ratios. We get different Character Error Rates for each level of SNR. By comparing the individual graphs of the different models, we are able to access and evaluate their robustness. Figure 2 shows the CER at different levels of SNR.

Confidential and Proprietary to CVDI

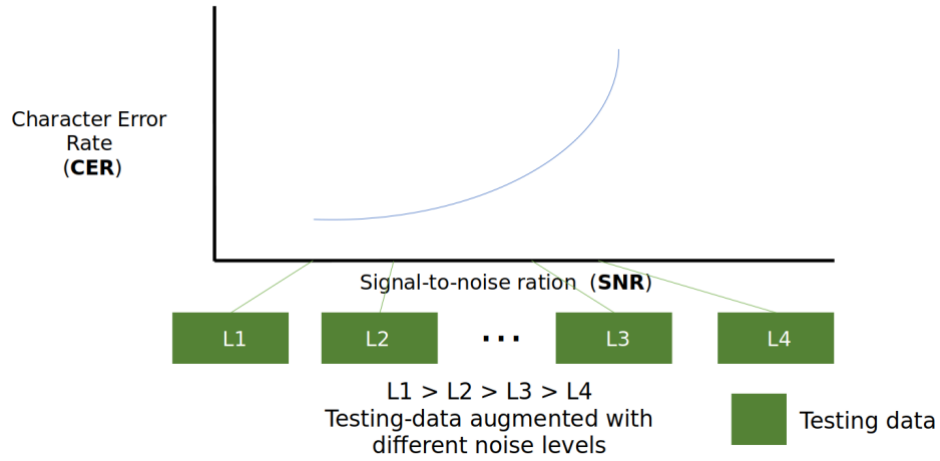

Figure 2. Graph of the CER at different levels of SNR

Augmentation Methods

We have studied and applied different data augmentation techniques such as changing the speaker acoustics using vocal tract length perturbation (VTLP), spectral shifts, and speech speed distortions. We also modified the acoustic environment, e.g. reverberation and (channel variability).

Training strategies

We have devised 3 different strategies to train our model, as shown in Figure 3. The first strategy is to start training with the clean training data and then proceed with the augmented data. The second strategy is to proceed the opposite way starting with the augmented data and then proceed to the clean data; while the last strategy is to shuffle the clean and augmented data during the training.

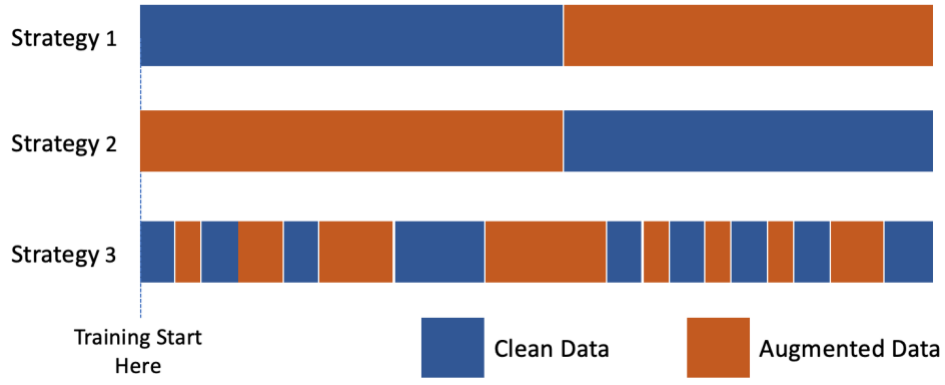

Figure 3. The 3 different Training Strategies

Confidential and Proprietary to CVDI

Results

The first two experiments establish the baseline, we have trained with an increasing number of utterances. As one would expect, training with more data would lead to a lower CER. But after crossing the 0db point, both models get similar CERs since they have been trained on clean data only.

The third model has been trained following training strategy 2. It used the augmented version of the data followed by the clean data. This latter model proved to be more robust compared to a model trained with clean data only. The difference gain below 20db is around 2-3%, because the model trained with only clean data would still perform well in inferring clean speech. Once the signal gets more corrupted, above 20db, the gap between the model trained with only clean data and the one trained with augmented data becomes almost 10% CER which is a 20% improvement. Check Appendix A for inference examples.

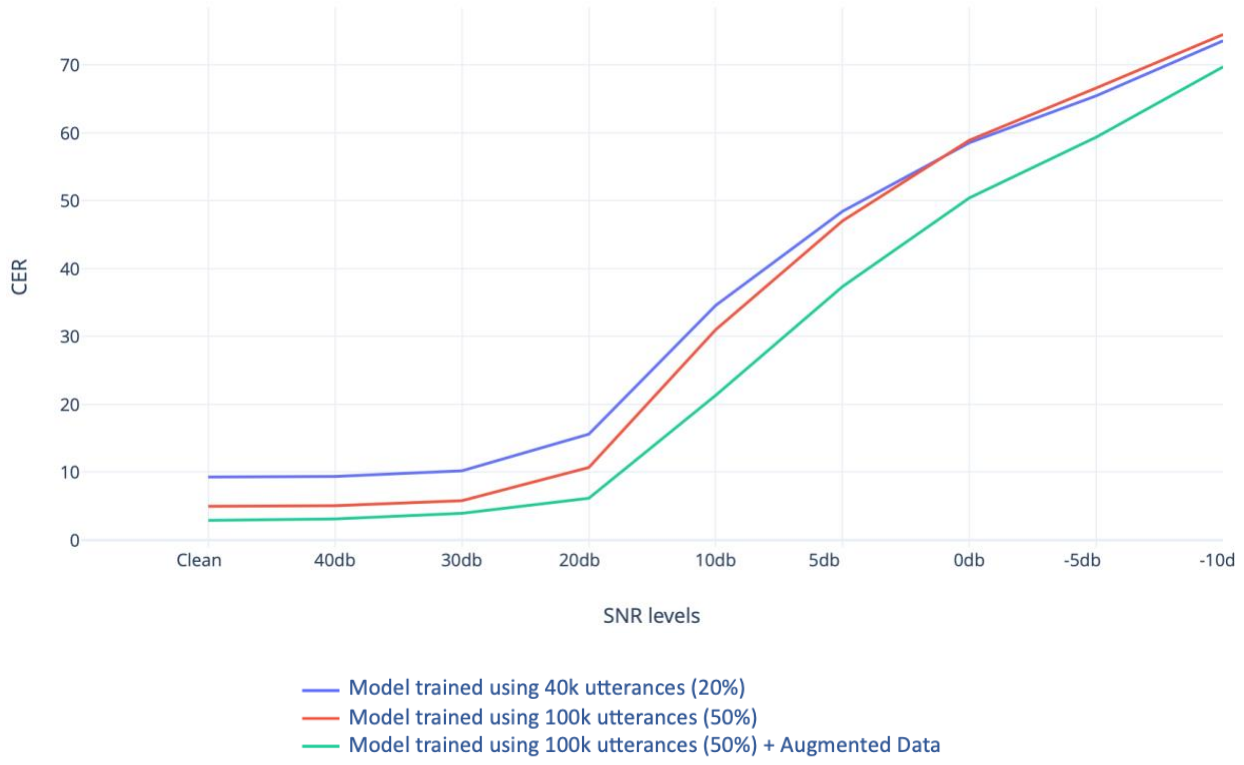

Figure 4. Models trained with different training folds.

Confidential and Proprietary to CVDI

	Trained with 20% clean Training Data	Trained with 50% clean Training Data	Trained with 50% clean plus the Augmented Data
Clean	9.283	4.963	**2.885**
40db	9.362	5.051	**3.102**
30db	10.194	5.789	**3.928**
20db	15.598	10.684	**6.148**
10db	34.578	30.997	**21.32**
5db	48.449	47.040	**37.357**
0db	58.561	58.928	**50.428**
-5db	65.479	66.630	**59.381**
-10db	73.590	74.518	**69.749**

Table 2. CER Results from 3 different training setups

Functionality of Innovation(s)

This training strategy helps building more robust automatic speech recognition system based on limited data resources. We utilize the augmented data alone to train the model to land into a good minima, and then we fine tune it with the clean version of that data.

Conclusions and Recommendations

Data augmentation is useful to train more robust ASR systems, especially when faced with limited datasets. These methods can be as simple as babble noise or speech speed distortions.

We have focused on improving the CER as a way to measure the capability of the model of classifying the given alphabet correctly. For better results on the Word Error Rate a language model can be used for a better inference result.

Impact and Uses/Benefits

The proposed approach can be used to train automatic speech recognition systems with limited datasets and computational resources for a specific use case. We have also defined a way to assess robustness of an ASR model.

Confidential and Proprietary to CVDI

Bibliography

Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., Casper, J., Catanzaro, B., Cheng, Q., Chen, G. and Chen, J., 2016, June. Deep speech 2: End-to-end speech recognition in english and mandarin. In International conference on machine learning (pp. 173-182).

Panayotov, V., Chen, G., Povey, D., & Khudanpur, S. (2015, April). Librispeech: an ASR corpus based on public domain audio books. In 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 5206-5210). IEEE.

Jaitly, Navdeep, and Geoffrey E. Hinton. "Vocal tract length perturbation (VTLP) improves speech recognition." In Proc. ICML Workshop on Deep Learning for Audio, Speech and Language, vol. 117. 2013.

Ko, Tom, Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur. "Audio augmentation for speech recognition." In Sixteenth Annual Conference of the International Speech Communication Association. 2015.

Cui, Xiaodong, Vaibhava Goel, and Brian Kingsbury. "Data augmentation for deep neural network acoustic modeling." IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP) 23, no. 9 (2015): 1469-1477.

Confidential and Proprietary to CVDI

Appendix A

Baseline
Inference of Clean Test Data

Ground Truth
I love thee with a love I seemed to lose with my lost saints I love thee with the breath smiles tears of all my life and if god choose I shall but love thee better after death

Inference with (~40% of the training dataset) model: (Prefect prediction)
I love thee with a love I seemed to lose with my lost saints I love thee with the breath smiles tears of all my life and if god choose I shall but love thee better after death

Inference with (~20% of the training dataset) model:
I loved thee with a love y seemed to loodse with my last sants I love thee with the breath smiles tear of awl my life and if dor thodes I shall but love the better after death

Inference with (~10% of the training dataset) model:
I lofe the with a love e semed to lo with my lost saints I lovedy with the breath smile t her of offr might a life and if go trus I shall but love thy better after death

Testing for robustness
Inference on Test Data with different SNR levels

Ground Truth
Though the discipline of the former parliamentary army was not contemptible a more exact plan was introduced and rigorously executed by these new commanders
Inferring clean speech

Inferring 30db SNR
Teyh the discipline of the former paliamentary army was not contemptible a more exact plan was introduced and rigorously executed by these new commanders

Inferring 20db SNR
Ogh the discipline of the fuller parlimentary army was not continltable a more exact plan was interodused and rigarously executed by these new commanders

Inferring 10db SNR
Euwthevizselin ant the foullor colinentay anly was not conpatale more that clan ith intyhuse and riitsly a lecutin by walli a camelan

Inferring 5db SNR
E hycithing in the holo calimitly aoet hor n qhatiga mor yar cran wie letkuthin hame loo

Inferring 0db SNR
O halitot the holo koonota erlluls ind famtwo u the planwitllily mlo

Confidential and Proprietary to CVDI