Does Speech Enhancement of Publicly Available Data Help Build Robust Speech Recognition Systems? (Student Abstract)

Bhavya Ghai,[1] **Buvana Ramanan,**[2] **Klaus Mueller**[1]

[1]Department of Computer Science , Stony Brook University, USA
[2]Nokia Bell Labs, Murray Hill, USA
{bghai, mueller}@cs.stonybrook.edu, buvana.ramanan@nokia-bell-labs.com

Abstract

Automatic speech recognition(ASR) systems play a key role in many commercial products including voice assistants. Typically, they require large amounts of high quality speech data for training which gives an undue advantage to large organizations which have tons of private data. We investigated if speech data obtained from publicly available sources can be further enhanced to train better speech recognition models. We begin with noisy/contaminated speech data, apply speech enhancement to produce 'cleaned' version and use both the versions to train the ASR model. We have found that using speech enhancement gives 9.5% better word error rate than training on just the original noisy data and 9% better than training on just the ground truth 'clean' data. It's performance is also comparable to the ideal case scenario when trained on noisy and it's ground truth 'clean' version.

Introduction

Automatic speech recognition(ASR) can be understood as a process to convert audio signal to text. ASR systems are a critical part of all voice assistants like siri, cortana, etc. Technology giants like Google, Amazon, etc. leverage tons of private data to build state-of-the-art ASR systems. This makes it really difficult for other players to reproduce similar performance. In this paper, we are trying to investigate if we can use publicly available data to train ASR systems which can compete with the state of the art. If true, it will empower startups, academics, etc. to build competent ASR systems. Publicly available speech data like youtube may be contaminated with ambient noise and background music which makes it difficult to be used for training ASR systems. Hence, we propose speech enhancement techniques to clean the noisy data first and then use the original and it's enhanced(cleaned) version to train ASR systems.

Speech enhancement(SE) is a well studied problem which aims to enhance audio quality by getting rid of contamination's such as white noise, background music, etc. Different GAN based models like SEGAN, FSEGAN, etc. have been shown to perform well for speech enhancement. In this work, we have used SEGAN (Pascual, Bonafonte, and Serra

2017) which operates at waveform level to remove noise from given noisy speech signal. SEGAN uses CNNs instead of RNNs for it's encoder and decoder modules which makes it faster. It operates end to end with raw audio signal so it's free of any assumptions made for feature extraction. Lastly, authors have also shared its code which makes it more reproducible. Hence, we chose SEGAN over other speech enhancement techniques.

There are different processing stages to go about building noise robust ASR systems. Deep learning approaches to build robust ASR systems can be classified into 3 groups i.e. front-end, back-end & joint front- and back-end techniques(Zhang et al. 2018). In the front-end setting, speech enhancement and recognition system are independent from each other. Noisy speech is first enhanced during pre-processing and then recognizer is trained on the enhanced speech. In the back-end setting, noisy speech is fed in as it is and the recognizer is optimized such that it outputs the correct corresponding text. Lastly, In joint front- and back-end setting, speech enhancement and recognizer are considered as a single block and trained end-to-end. In this work, we have focused on the front-end and back-end approach as shown in Fig. 1.
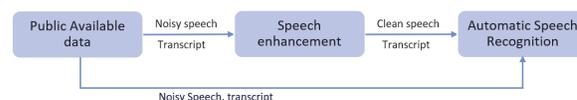

Figure 1: Our Approach: We use publicly available noisy data and it's cleaned version to train ASR model.

One of the most popular approach for back-end setting is multi-condition training. Multi-condition training is a technique which helps make more robust recognition system by training on multiple acoustic variants of the training dataset. In our case, we propose to use publicly available noisy speech along with it's cleaned variant(via SE) for building ASR systems.

Dataset

Existing datasets for speech enhancement containing ground truth-clean and contaminated audio pairs are pretty limited

in size. ASR systems trained on such datasets might not generalize to different real world conditions. Hence, we decided to curate our own speech enhancement dataset. For ground truth-clean speech, we used LibriSpeech dataset (Panayotov et al. 2015) which is derived from public domain audiobooks. This dataset is fairly big(~460hrs) and its corresponding transcript is also available which makes it suitable for ASR training. Next, to get its contaminated version, we distorted the audio by adding diverse set of background music and ambient noises to simulate different real world conditions. For ambient noise, we used popular datasets like Urbansound, ESC50 along with Youtube, specifically tracks of background noises in train, traffic, restaurant, rain, etc. For background music, we used Youtube to extract movie theme songs and instrumental music belonging to different genres like Latin, African, Heavy metal, etc. This resulted in ~205hrs of contaminated/noisy mixture for which we also possess ground-truth clean variant and the transcript.

Experiments

First, we tried to investigate if training with noisy and it's ground truth-clean variant really helps. We trained DeepSpeech model (Hannun et al. 2014) as the recognizer on 100hrs of clean, noisy mixture and clean+noisy mixture (clean refers to ground truth-clean data). We tested the model on 5hrs of clean and noisy test dataset. For evaluation, we have used the de-facto standard for ASR systems which is word error rate(WER). WER is the percentage of words mis-recognized by the ASR system(lower the better). To compare our results, we have considered 3 different cases i.e., real world scenario(noisy), ideal case(noisy+clean) and our solution(noisy+enhanced). In the real world scenario, we can gather noisy data from public sources. So we trained ASR system just on noisy data. For the best case scenario, we trained DeepSpeech with noisy dataset and it's clean version. Lastly, we implemented our approach. We first processed the noisy dataset with pretrained SEGAN to get enhanced dataset. Thereafter, we trained the DeepSpeech model with noisy dataset and it's enhanced version. If our Speech enhancement model works really well, only then we might achieve similar performance as the ideal case. First two cases represent back-end approach because there is no preprocessing involved. The model is left to decide what is noise and what is not. Our approach is a hybrid of front-end and back-end approach because we clean the speech first and feed noisy and it's enhanced version for training.

Results

As shown in Fig. 2, DeepSpeech model trained on clean data performs well on clean test set but lags on noisy test set. Similarly, when trained on noisy data, DeepSpeech model performs better on noisy test set but lags on clean test set. Finally, DeepSpeech model trained on clean+noisy mixture outperforms other two cases on both clean and noisy test set. So clearly, training with noisy and clean version helps.

Since we mightn't always have the clean version of publicly available data, we replaced clean speech with enhanced speech. Noisy speech combined with it's enhanced speech
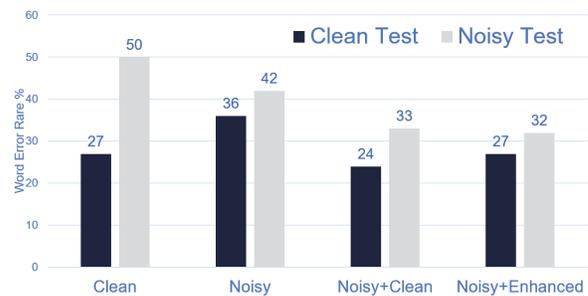

Figure 2: Word error rate of DeepSpeech model when trained and tested on different datasets. X-axis represents different kinds of training data i.e. clean, noisy, etc. Dark colored bars represent word error rate when tested on clean data while light colored bars represents noisy test data

by SEGAN performed significantly well for ASR systems. We observed 9.5% mean reduction in WER on noisy(42% vs 32%) and clean(36% vs 27%) test data when compared with real world scenario i.e. training on just noisy speech. We observed that speech cleaned with SEGAN performed at par with the ideal case scenario for noisy test dataset(32% vs 33%). On the clean test set, it's error rate is a little higher than ideal case(27% vs 24%). This might be attributed to the artifacts introduced by speech enhancement model on clean speech. In conclusion, this work is a proof of concept that found data treated with some speech enhancement model helps ASR become more robust and accurate.

Conclusion & Future Work

Our work shows that publicly available data together with Speech enhancement models can be leveraged to build robust ASR systems. Next, we intend to test our approach with other SE models like FSEGAN, Wave-u-net, etc. It will also be interesting to test how back-end approach compares with end-to-end approach. Overall, we believe this work will motivate larger research on building state of the art ASR systems from public available/found data.

References

Hannun, A.; Case, C.; Casper, J.; Catanzaro, B.; Diamos, G.; Elsen, E.; Prenger, R.; Satheesh, S.; Sengupta, S.; Coates, A.; et al. 2014. Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567*.

Panayotov, V.; Chen, G.; Povey, D.; and Khudanpur, S. 2015. Librispeech: an asr corpus based on public domain audio books. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, 5206–5210. IEEE.

Pascual, S.; Bonafonte, A.; and Serra, J. 2017. Segan: Speech enhancement generative adversarial network. *arXiv preprint arXiv:1703.09452*.

Zhang, Z.; Geiger, J.; Pohjalainen, J.; Mousa, A. E.-D.; Jin, W.; and Schuller, B. 2018. Deep learning for environmentally robust speech recognition: An overview of recent developments. *ACM Transactions on Intelligent Systems and Technology (TIST)* 9(5):49.