Does Speech enhancement of publicly available data help build robust Speech Recognition Systems?

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 clean speech data for training which gives an undue advantage to large organizations which have tons of private data. In this paper, we have first curated a fairly big dataset using publicly available data sources. Thereafter, we tried to investigate if we can use publicly available noisy data to train robust ASR systems. We have used speech enhancement to clean the noisy data first and then used it together with its cleaned version to train ASR systems. We have found that using speech enhancement gives 9.5% better word error rate than training on just noisy data and 9% better than training on just clean data. It's performance is also comparable to the ideal case scenario when trained on noisy and its 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 its 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 contaminations 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 its encoder & decoder modules which makes it faster. It operates end to end with raw audio signal so its 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 architectures to go about speech enhancement for 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 preprocessing and then recognizer is trained on the enhanced speech. In the back-end setting, noisy & enhanced speech are used together to train recognizer. Lastly, In joint front- and back end setting, speech enhancement & recognizer are considered as a single block and trained end-to-end. In this work, we have focused on the back-end technique as show in Fig. 1.

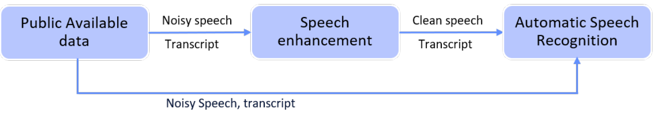


Figure 1: Our Approach: We use publicly avialable noisy data and its cleaned version to train ASR model.

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 its cleaned variant(via SE) for building ASR systems.

Dataset

Existing datasets for speech enhancement are pretty limited in size. ASR systems trained on such datasets mightn't generalize to different real world conditions. Hence, we decided to curate our own dataset. For 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, we used diverse set of background music & ambient noises to simulate different real world conditions. For ambient noise, we used popular datasets like Urbansound, ESC50 along with youtube. From youtube, we cherry picked videos reflecting 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, Native American, Japanese, Indian, African, Heavy metal, etc. Lastly, we added ambient noise and background music to the clean speech. This resulted in ∼205hrs of noisy mixture for which we possess its clean variant along with the transcript.

Experiments

First, we tried to investigate if training with noisy & its clean variant really helps. We trained deep speech model (Hannun et al. 2014) (recognizer) on 100hrs of clean, noisy mixture and clean+noisy mixture. We tested the model on 5hrs of clean and noisy 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).
As shown in Fig. 2, deep speech model trained on clean data performs well on clean test set but lags on noisy test set. Similarly, when trained on noisy data, deep speech model performs better on noisy test set but lags on clean test test. Finally, deep speech model trained on clean+noisy mixture outperforms other two cases on both clean and noisy test set. So clearly, training with noisy & clean version helps.

To compare our results, we have considered 3 different cases i.e., real world scenario (noisy), ideal case (noisy+clean), 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 its clean version. If our Speech enhancement model works really well, only then we might achieve similar performance. 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 its enhanced version. First two cases represent back end approach because we there is no preprocessing involved. The model is left to decide what is noise and what is not. Our approach falls under the the front end approach because we clean the speech first before training.

Results

In a multi-style ASR training, noisy speech together with its cleaned version can significantly reduce word error rate. Since we don't have clean version of publicly available data, we replaced clean speech with enhanced speech. Noisy speech combined with its enhanced speech by SEGAN performed significantly well for ASR systems. We observed 9.5% reduction in WER when compared with noisy speech. We observed that speech cleaned with segan performed at par with the ideal case scenario for noisy test dataset. On the clean test set, its error rate is a little higher than ideal case. 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 & accurate.

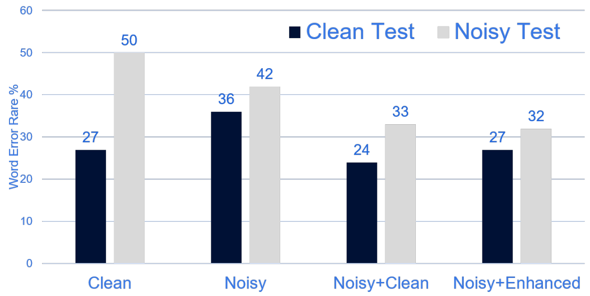


Figure 2: Word error rate of DeepSpeech model when trained and tested on different datasets. X-axis represents different kinds of training data i.e. clean, noisy, etc. Dark colored bars represent word error rate when tested on clean data while light colored bars represents Noisy test data

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.

Dataset

One of the contributions of this paper is to build a large dataset which reflects real world environments using public sources. For clean speech, we used LibriVox dataset as cited in the paper. For background noise, we used two publicly available dataset i.e. UrbanSound & ESC50. Furthermore, we used a bunch of videos from youtube. To enhance reproducibility, its critical to provide links for all these sources. Following are the links for all sources along with their duration:

Urbansound https://urbansounddataset.weebly.com/urbansound8k.html (5.5 hours)

ESC 50 https://github.com/karoldvl/ESC-50 (2.75 hours)

Youtube background noise (56 hours)

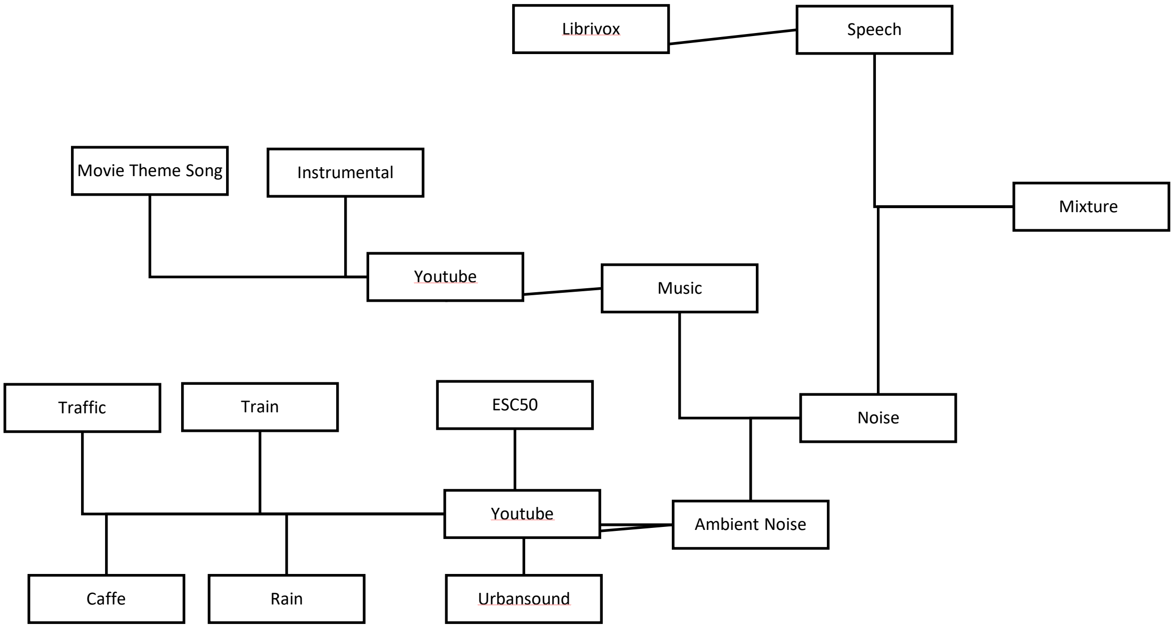
- HighWay Sound (~8hrs) https://www.youtube.com/watch?v=AVIDrl4ZNJ4
- Restaurant Sound (~8hr) https://www.youtube.com/watch?v=SYpJDOn1myo
- Car driving in the rain (~2hr) https://www.youtube.com/watch?v=cDNETgwyQSI
- Airport Sounds (~8hr) https://www.youtube.com/watch?v=FhBWRhtZL28
- Coffee Shop background noise (~47min) https://www.youtube.com/watch?v=Bp9qoIUFRUw
- Train in rain (~2hr) https://www.youtube.com/watch?v=FhYaXj91juE
- City Traffic(~8hr) https://www.youtube.com/watch?v=8s5H76F3SIs
- Shopping Mall Sound (~8hr) https://www.youtube.com/watch?v=kednn_OXJ4Q
- 10 hours of Relaxing Restaurant Ambient Background Noise - Crowd Sound Effects - Cafe Soundtrack https://www.youtube.com/watch?v=Ay4c4yAB-so
- Forest and Nature Sounds 10 Hours https://www.youtube.com/watch?v=OdIJ2x3nxzQ
- CITY SOUNDS, GHETTO SOUNDS WHITE NOISE, SOUNDS OF THE CITY, CITY SOUNDS WHITE NOISE FOR SLEEP (8hrs) https://www.youtube.com/watch?v=Wc3XjE-3zYE&t=14037s

Youtube Ambient Music (44 hours)

- Indian Meditation Music for Positive Energy Flute Music Indian Krishna (3:19:34) https://www.youtube.com/watch?v=zZFqDWLsdkA
- 4 Hours of The Best Epic Inspirational Music for Studying/Working (4:13:05) https://www.youtube.com/watch?v=0tuK0sk_D1M
- Native American Flute Music: Meditation Music for Shamanic Astral Projection, Healing Music (9:09:08) https://www.youtube.com/watch?v=ST56ATKfgfs
- Heavy metal hard rock music instrumental compilation (1:40:35) https://www.youtube.com/watch?v=BxDiQhNO780
- Folk Music Instrumental 10 Hours (10:00:00) https://www.youtube.com/watch?v=A0ggT0eewsY
- 3 HOURS of the Best Traditional Japanese Music - Relaxing Music for Stress Relief and Healing (2:59:00) https://www.youtube.com/watch?v=pPFabRaQI-0
- 2 Hours Of Instrumental Latin Music - Salsa, Tango, Bachata, Rumba (2:10:00) https://www.youtube.com/watch?v=1ZENiMeCT84
- 1 Hour of African Folk Music Instrumental | Marimba, Kalimba, & Drums (1:06:28) https://www.youtube.com/watch?v=9b81mWYIyTo

- 8 Hours of Epic Inspirational Music for Studying/Working - Best Movie & TV Soundtracks (8:03:25) https://www.youtube.com/watch?v=hPxW-oyVFig
- Inspiring Movie Soundtracks (1:25:08) https://www.youtube.com/watch?v=E7kRQAy9tho
- Wonderful movie soundtracks (1:23:25) https://www.youtube.com/watch?v=c9V3FBJ8FoA

Dataset is created by adding all these noise sources to clean speech. Following is a visual representation of how the final mixture is created


Librivox
Speech
Movie Theme Song
Instrumental
Mixture
Youtube
Music
Traffic
Train
ESC50
Noise
Youtube
Ambient Noise
Caffe
Rain
Urbansound