Development and Evaluation of a Novel Speech Recognition System Using Listen, Attend, and Spell and a DRNN Denoiser

Bradley He

Abstract:

Speech denoising has become an increasingly important part of automatic speech recognition systems (ASR). Use of neural networks for denoising audio has been gaining traction, but has not been widely tested working in tandem with ASR models. This project aimed to fill this gap in knowledge by measuring the performance of a speech denoising model by looking at the performance of a recently-developed neural network speech recognition system. Testing denoising models by the performance of an ASR model shows how the denoised sound would be understood in a more practical scenario. Denoised speech output was tested on the ASR model Listen, Attend, and Spell (LAS). It will measure the effectiveness of the denoising model by the phoneme error rate of the LAS model compared to the correct phoneme sequence. Speech data will be obtained from the TIMIT data set. Speech samples were mixed with background noise at sound to noise ratios ranging from -20 dB to 30 dB at 5 dB intervals. Noisy speech samples were denoised using a DRNN denoiser, then tested using a LAS model. The phoneme error rate of the model's prediction on the denoised voice samples compared to the correct phoneme sequence was used as the indicator of the model's performance. This project revealed that denoisers can work in tandem with ASR models like LAS and do improve performance in noise-dominated speech samples, but can often worsen performance when there is little to no noise present.

Introduction:

Automatic speech recognition (ASR) is the task of converting spoken words into transcribed text. In the modern day, ASR has become integral to the lives of many people through the usage of voice assistants and accessibility technology, where accurate transcription of voice to text is important. A variety of models and techniques have been used to perform this task of speech-to-text translation, but the usage of deep neural networks (DNNs) in tandem with Hidden Markov Models (HMMs) to create a hybrid DNN-HMM system has emerged as the "state-of-the-art" front-runner in ASR models in the 2010s (Jaitly et al., 2012). Such a hybrid system is necessary due to the nature of the task of speech recognition and transcription: DNNs, while excellent at tasks such as classification, which usually have fixed-length inputs and outputs, are not suited for doing abstract tasks such as speech recognition, which usually have variable-length inputs and outputs (Chan et al., 2015). HMMs are able to resolve this issue since they are sequential models, which allow for variable-length inputs and outputs to be processed (Leonart et al., 1966). Thus, the DNN-HMM system is able to take in variable-length audio samples and output variable-length text transcriptions. The DNN-HMM system, however, has several drawbacks. In such a system, the numerous components — for example, the acoustic model, pronunciation model, or the language model — are all trained separately and with a different objective, which can cause disjointed and uncoordinated learning that degrades performance (Chan et al., 2015). Additionally, the HMM component of the system lacks adaptability and flexibility since it must be manually generated (Graves et al., 2006). Attempts to resolve these issues have resulted in the development of "end-to-end" or "sequence-to-sequence" ASR models, in which the audio sample is directly converted to a text transcript in one step,

without the usage of HMMs. Listen, Attend, and Spell (LAS), developed in 2015, is one such end-to-end ASR model, which uses an entirely neural network-based framework. To deal with the variable-length input problem, it utilizes an encoder and a decoder recurrent neural network (RNN), with an attention mechanism between the two. The encoder RNN takes in a variable-length input and outputs a fixed-length output, which is modified by the attention mechanism and inputted back into the decoder RNN. The decoder RNN is then able to output a variable-length output. Utilizing a purely neural network-based framework allows for more adaptability and flexibility compared to HMMs since the model is generated "naturally" through self-learning rather than "artificially" through human interference (Graves et al., 2006). As a result, LAS is able to perform at the level of conventional, state-of-the-art DNN-HMM systems (Chan et al., 2015), and can even outperform DNN-HMM systems with certain modifications such as label smoothing (Chiu et al., 2018) and data augmentation (Park et al., 2019).

The model used for speech recognition is only one part of the speech recognition process. One can also use a noise removing model, or a "denoiser", to improve the quality of the voice input. Similar to ASR models, many denoising models use "conventional" methods that involve hand-engineered components, such as filters and transforms, to remove unwanted noise (Maas et al., 2012). Recently, neural network-based denoisers have overtaken such conventional models, as DRNN-based models have shown to perform significantly better than non-neural network based models (Huang et al., 2015). The metric of determining how well a denoiser performs, however, is not as straightforward as it may seem. Research in the past has evaluated denoiser performance through direct comparison of the denoiser output with the ideal audio output: for example, Huang et al., 2015 and Grais et al., 2017 utilize metrics such as SDR, which measures

the difference between the denoiser output and the noisy input, SAR, which measures unwanted artifacts added during the denoising process, and SIR, which measures the amount of noise present after the denoising process. Although these metrics give insight as to how well the denoiser removes the noise from a noisy input, they fail to indicate how well the denoiser would work in a real-life scenario, where the denoiser would be coupled with an ASR model. Some research has been conducted in an attempt to resolve this problem: Maas et al., 2012 evaluates an RNN denoiser's output using a conventional ASR model and uses the ASR model's performance as a metric for denoiser performance. However, little research has been done into the interactions between denoisers and more recent and relevant ASR models.

This project is an attempt to resolve this gap in knowledge by evaluating denoiser performance using newer ASR models, specifically LAS. It investigates the interactions between the DRNN denoiser described in Huang et al., 2015 and a modified ASR model based on the LAS model described in Chan et al., 2016.

Methods:

TIMIT Dataset:

There are many publicly available voice datasets, but the 1988 TIMIT dataset, a collection of over 5000 audio files containing spoken sentences (Lopes et al., 2011), was chosen for testing due to its usage of phonemes, distinct phonetic segments that make up words, and its relative ease of usage. Phoneme-based datasets, compared to word-based datasets, are more desirable since they are not inherently restricted by a limited vocabulary (Lopes et al., 2011). In the TIMIT dataset, every voice sample is paired with its corresponding text transcription in

phoneme form. The 61 phoneme dictionary of the TIMIT dataset is usually considered too narrow of a classification for ASR models to accurately differentiate between, so all 61 phonemes were collapsed into a set of 39 phonemes that were more broadly classified (Lee et al., 1989).

Voice samples in the TIMIT dataset are separated into two sets: a training set, and a testing set. The training set consisted of 3696 spoken sentences amounting to 3.14 hours of speech time, and the testing set consisted of 1344 spoken sentences amounting to 0.81 hours of speech time (Lopes et al., 2011). Both of the sets were used in this project for their respective purposes.

Noising and Denoising Process:

The denoiser used to denoise artificial noisy speech samples was the DRNN denoiser described in Huang et al., 2015, which was publicly available on Github. MATLAB R2015a was used to run the denoiser.

Simulating real-life usage of a denoiser-ASR model tandem system requires generating artificially noisy voice samples to test and evaluate (Maas et al., 2013). To create one such noisy voice sample, various noise samples were obtained from Huang et al., 2015. These consisted of traffic, talking, construction noise, and subway noise, among others. A noisy voice sample was generated by mixing a randomly selected noise with a given voice sample. To help standardize the intensity of noise compared to the voice when mixing noise and voice together, a metric called “target to interference ratio” (TIR) was used. TIR measures how much louder the voice is compared to the noise in decibels (dBs): for example, a TIR of 15 indicates that the voice is 15

dB louder than the noise, and a TIR of -10 indicates that the voice is 10 dB quieter than the noise. After generating a noisy voice sample, the DRNN denoiser could then be used to denoise the voice sample. The resulting output of the DRNN denoiser was called the denoised voice sample.

22 individual testing datasets were generated, with noisy testing datasets and denoised testing datasets each comprising half of the total. Noisy testing datasets were generated by taking the unmodified TIMIT testing dataset and mixing noise with each voice sample in it at a specific TIR ranging from -20 dB to 30 dB in intervals of 5 dB (i.e. -20 dB, -15 dB, -10 dB, … 25 dB, 30 dB) (Maas et al., 2013), generating 11 TIR-specific noisy testing datasets. Denoised testing datasets were generated by taking each noisy testing dataset and running every voice sample through the DRNN denoiser, which generated 11 TIR-specific denoised testing datasets for a total of 22 testing datasets.

An additional "denoised" training dataset was generated as well. In a similar process to the generation of the denoised testing datasets, all voice samples in the unmodified training dataset were mixed with noise at a TIR of 0 dB and denoised, which generated a denoised training dataset. This denoised training dataset, along with the unmodified training dataset, were used to train two separate LAS models, which is discussed in the next section.

Model Generation and Training:

The model used to evaluate speech samples is the aforementioned LAS model with label smoothing implemented (Chiu et al., 2018), which allows for greater stability and accuracy

during training. A Pytorch implementation of the model was made freely available by Alexander H. Liu on Github.

Two LAS models were generated for use in this experiment. The first, "clean" model was trained on the unmodified training dataset, while the second, "denoised" model was trained on the denoised training dataset. Both datasets were divided into a 95% to 5% split of training data to validation data, where both models were trained on the training data until their performance on the validation data did not improve (Palaz et al., 2013). Both the clean and the denoised model took around 5 days for the training process to finish.

Model Evaluation of Speech Samples:

After evaluating a given voice sample and outputting its prediction of the text transcript in a phoneme sequence, the model's performance is measured using a metric called phoneme error rate (PER), which is the Levenshtein distance between the predicted phoneme sequence and the actual phoneme sequence (Lopes et al., 2010). A lower PER indicates better performance of the model, and a higher PER indicates worse performance of the model. An LAS model trained on the unmodified training set and tested on the unmodified testing set achieves an average PER of 26%, which can be seen as the optimal performance of any LAS model.

All 22 testing datasets were evaluated on both the "clean" model and the "denoised" model to generate a total of 44 PER values. Thus, 11 PER values representing TIRs from -30 dB to 20 dB were generated for each combination of clean or denoised model and noisy or denoised voice samples. The clean model evaluated with the noisy testing dataset was viewed as the control group of the experiment, since it would represent the performance of a normal LAS

model in a noisy environment. Likewise, the usage of the denoised testing dataset and the denoised LAS model was viewed as independent variables.

Results/Discussion:

Two LAS models, one trained on the unmodified TIMIT training set and one trained on a denoised training set, were used to evaluate noisy and denoised testing datasets at TIRs ranging from -20 dB to 30 dB. All four groups showed a consistent decline in PER with an increase in TIR, which was expected since higher TIRs indicated lesser amounts of noise in the input. The clean model evaluated with noisy input, or the control group, consistently performed worse than the three experimental groups between TIRs of -20 dB and 15 dB, albeit still decreasing significantly. This is expected behavior as well, since denoising noisy input would reduce the amount of noise present in the input and allow for a more concise transcription. However, between TIRs of 15 dB and 30 dB, the control group unexpectedly performed better than the three experimental groups (Figure 1). In fact, at a TIR of 30 dB, the control group is only 2% in PER above its optimal performance, whereas the experimental groups are well over 10% in PER above the optimal performance. This could be attributed to the DRNN denoiser targeting and degrading the quality of the actual voice of the voice sample due to the lack of any significant noise at high TIRs, which would make denoised testing datasets harder to interpret. Conversely, the control group would have a much easier time interpreting high-TIR voice samples since the noise at such TIRs is negligible.

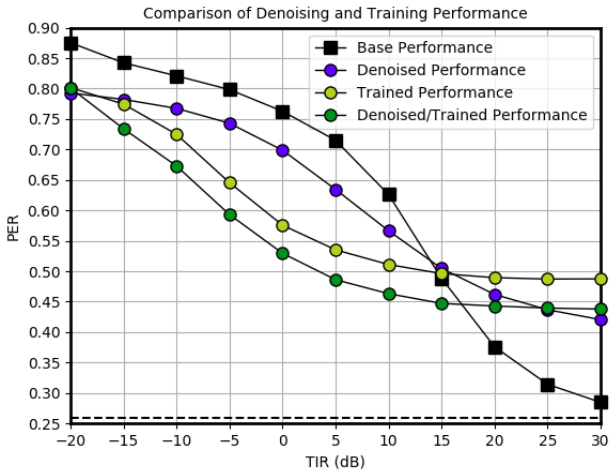

Figure 1: Comparison of the control group and the three experimental groups, with phoneme error rate (PER) plotted against target-to-interference ratio (TIR) in dB. Optimal performance of the LAS model is indicated by the dashed line.

Direct comparisons between the three experimental groups and the control group yield more detailed findings. By subtracting each experimental group's PER across all TIRs from the control group's PER, relative improvement values were obtained. The relative improvement of voice sample denoising corresponds to the findings from Figure 1, with performance improvements only present up to 10 dB, after which higher TIRs result in worsened performance (Figure 2a). Model training on the denoised dataset, however, shows significantly better performance improvement than voice sample denoising, with all TIR values from -20 dB to 15 dB showing greater relative improvement in model training than in voice sample denoising (Figure 2b). Two factors are likely behind this behavior. The first factor is that the DRNN denoiser may not always fully remove the noise in noisy speech samples, particularly at low TIRs (Huang et al., 2015). As a result, there is still significant residual noise after denoising, which diminishes the effectiveness of denoising the speech samples. The second factor is a direct consequence of the first factor: since the denoiser failed to completely remove the noise from the

speech samples, the speech samples in the denoised training dataset likely contained some noise. As a result, the deniosed LAS model, through training, was able to interpret the voice sample through noise more accurately. The performance improvements when both voice sample denoising and model training on the denoised training dataset were significantly greater than either of the two techniques alone (Figure 2c); for example, at a TIR of 15 dB, while the two separate approaches both saw worsened performances, the combined approach was able to show performance improvement. However, the combined approach was still unable to counteract the worsening performance at the higher end of TIRs that was present in the individual techniques.

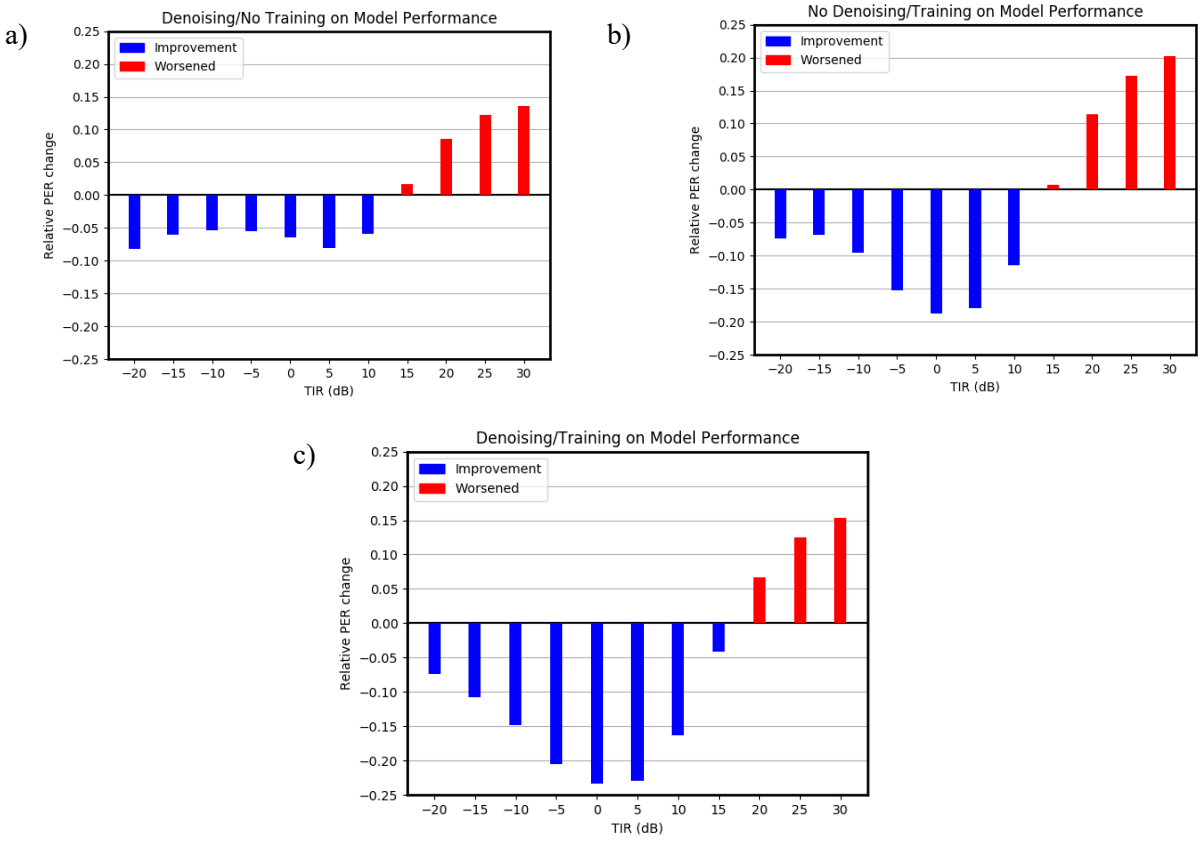

Figure 2: Performance improvements with usage of voice sample denoising (a), training on the denoised training dataset (b), and both (c), with relative PER change plotted against target-to-interference ratio (TIR) in dB. Blue bars indicate relative improvement, whereas red bars indicate relative worsening.

Conclusion:

This study is the first to explore the interactions between the LAS model and DRNN denoisers. Although previous studies have successfully created tandem ASR model-denoiser systems using conventional ASR models such as HMMs, little research has been done regarding newer ASR models and denoisers. Using the TIMIT dataset, this study reveals that denoisers can successfully be used in tandem with LAS through either voice sample denoising or training data denoising, but with certain drawbacks. At higher TIRs, where noise is negligible to the voice sample, denoising apparently damages the voice sample to a certain extent such that an LAS model cannot interpret it as well as it could. In more realistic noise environments, however, the denoiser is able to significantly improve the performance of LAS, reducing the PER by as much as 33% in some cases.

Although these findings provide valuable insight into LAS-denoiser interactions, several aspects of this study limit the extent to which definitive results can be drawn. The first issue is the dataset used. LAS models are not well suited for training with small datasets such as TIMIT (Chan et al., 2016), and the model performance was likely to have been negatively impacted as a result. Usage of larger, more recent datasets such as LibriSpeech may provide more accurate and reliable results. The second issue is the lack of more models trained on denoised training sets at different TIRs. In this study, only two models were generated and utilized: a clean model trained on the unmodified training set, and a denoised model trained on a denoised training set at a TIR of 0 dB. Denoised models trained at TIRs lower or higher than 0 dB would give more insight into how training LAS on denoised voice samples affects model performance, and might perhaps provide a solution to the performance degradation seen at high TIRs.

Bibliography:

Baum, Leonard E., and Ted Petrie. (1966). Statistical Inference for Probabilistic Functions of Finite State Markov Chains. *The Annals of Mathematical Statistics*, vol. 37, no. 6, 1966, pp. 1554–1563.

Chan, W., Jaitly, N., Le, Q., & Vinyals, O. (2016). Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Chiu, C.-C., Sainath, T. N., Wu, Y., Prabhavalkar, R., Nguyen, P., Chen, Z., Bacchiani, M. (2018). State-of-the-Art Speech Recognition with Sequence-to-Sequence Models. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP).

Grais, Emad M., & Mark D. Plumbley. (2017). Single Channel Audio Source Separation Using Convolutional Denoising Autoencoders. *2017 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*.

Graves, A., Fernández, S., Gomez, F., & Schmidhuber, J. (2006). Connectionist temporal classification. Proceedings of the 23rd International Conference on Machine Learning - ICML 06.

Huang, P., Kim, M., Hasegawa-Johnson, M., & Smaragdis, P. (2015). Joint Optimization of Masks and Deep Recurrent Neural Networks for Monaural Source Separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(12), 2136-2147.

Jaitly, N., Nguyen, P., W. Senior, A., Kim, M., & Vanhoucke, V. (2012). Application of Pretrained Deep Neural Networks to Large Vocabulary Speech Recognition. *INTERSPEECH 2012.*

Lee, K.-F., Hon, H.-W., Hwang, M.-Y., Mahajan, S., & Reddy, R. (1989). The SPHINX speech recognition system. *International Conference on Acoustics, Speech, and Signal Processing.*

Lopes, C., & Perdigao, F. (2011). Phoneme Recognition on the TIMIT Database. *Speech Technologies*.

Maas, Andrew & Le, Quoc & O'Neil, Tyler & Vinyals, Oriol & Nguyen, Patrick & Ng, Andrew. (2012). Recurrent Neural Networks for Noise Reduction in Robust ASR. 13th Annual Conference of the International Speech Communication Association 2012, *INTERSPEECH 2012.* 1.

Palaz, D., Collobert, R., Perdigao, F., Salamin, H., & Magimai-Doss, M. (2013). End-to-end Phoneme Sequence Recognition using Convolutional Neural Networks. *ArXiv, abs/1312.2137.*

Park, D. S., Chan, W., Zhang, Y., Chiu, C.-C., Zoph, B., Cubuk, E. D., & Le, Q. V. (2019). SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. *Interspeech 2019.*