Speech Recognition in Noisy Environment Using Separate Speech Enhancement Modules

Zinan Zhuang
EID: zz6486
zinan@utexas.edu

Linyue Chen
EID: lc39897
linyuechen@utexas.edu

Abstract

A popular benchmark in automatic speech recognition tasks has been the word error rate on clean LibriSpeech data. In the real application, however, audios are usually mixed with diverse noises where the model trained on clean speech will be significantly impacted. We target the speech recognition problem in noisy scenario. We train a separate speech enhancement module before a pretrained ASR model and fine-tune them separately on a noisy dataset. The pretrained model with a low WER on the clean LibriSpeech performs poorly on our noisy dataset. With the speech enhancement module, the same model achieves a relative 40% error reduction; fine-tuning the ASR model yields a relative 55% error reduction. Finally, we test the performance by directly fine-tune the ASR model with noisy dataset; while it can achieve a performance similar to including a speech module, it performs significantly worse than fine-tuning on enhanced audio.

1 Introduction

Automatic speech recognition model iterates fast. Newer models are constantly developed and improved the state-of-the-art performance. A common test benchmark used is the LibriSpeech test-clean. ASR models that work well on clean audios, however, are susceptible to background noises, and the result would be the significantly degraded performance. The main reason is the mismatch between clean training features and noisy test features. A viable solution, as shown in this study, is to preprocess the noisy input by removing the noise interference before passing it into the ASR model using neural network-based speech enhancement module.

Speech enhancement in general learns to remove noise from the source audio. There are studies where speech enhancement modules are used to map distorted speech to clean speech (Narayanan and Wang, 2014) and (Chen et al., 2017). Two common approaches in single-channel speech enhancement are mask learning and feature learning: in the former method, the model creates mask based on noisy features, and the mask is then applied to the noisy audio to mask the noise interference. In the latter method, models works like seq2seq which map noisy speech directly to clean speech. Study (Narayanan and Wang, 2014) showed that due to the wide range for the values of masks, the mask-based learning converges faster and usually outperform the direct feature mapping. In our study, we use mask-based learning due to its quick convergence.

Our experiments are three folds. We first use a pretrained ASR model trained on LibriSpeech clean dataset as our baseline. The model performs well on clean dataset, but suffers significant performance loss on our noisy dataset. Then we train an LSTM-based speech enhancement module and insert it before the pretrained ASR model. Unlike (Narayanan and Wang, 2014), our enhancement module can work individually for enhancing audio. Finally we fine-tune the ASR model using SE generated audio based on the assumption that SE module could introduce its own characteristics into the audios. We also make a comparison between fine-tuning the ASR with and without SE module. The result shows that including a SE module significantly improve the ASR performance in all cases.

1.1 Related Work

The LibriMix dataset (Cosentino et al., 2020) consists of mixed LibriSpeech utterance and WHAM noises to simulate noisy background. We configure its script to generate single-channel noisy audios for our study.

Both (Chen et al., 2017) and (Strake et al., 2020) used two separate DNN for noise suppression and speech restoration. We used only one model to handle noise suppression with a much shallow depth, and the audio are synthesized mathematically.

2 Methods

2.1 Pretrained ASR

We are interested in re-using a well performing ASR for two reasons. First, it saves us a lot of time to train a strong ASR model from scratch given the major focus of is the speech enhancement module. Second, we want to compare the performance of fine-tuning with and without SE module. The pretrained model provides a high upper bound for fine-tuning.

2.2 Speech Enhancement Module

The SE module takes noisy wavform as input and extracts its magnitude features. The mask are created from noisy magnitude features and applied back to noisy features as enhanced features. Finally, the enhanced waveform are re-synthesized by calculating Inverse Short-term from enhanced magnitudes and noisy phases.

We use similar noise suppression architecture as in (Chen et al., 2017) and (Strake et al., 2020) where LSTM layers are sandwiched by fully connected layers. However, we use bi-direction LSTM instead of single LSTM layer; we also reduce the number of linear layers by keeping only one on both ends.

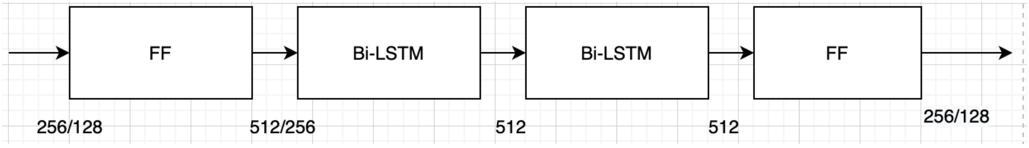

Figure 1: SE model architecture with input/output dimensions listed between layers

Figure 1 shows the structure of one LSTM block. We only use a total of two LSTM blocks and an extra ReLU activation as last layer to avoid negative output. The first LSTM block accepts 256-dimension magnitude features and output 128-dimension projections; the second block accepts the projections and output 256-dimension predicted magnitude features. The model capacity is traded for simplicity due to much less computation and faster convergence. The experiment section later shows the effectiveness of our SE module that even with concise architecture and mathematical synthesizer, there has been a huge word error reduction.

2.3 Metrics

2.3.1 Signal Approximation Loss

For training speech enhancement module, we use signal approximation which minimizes the clean signal spec and predicted signal spec. We implement it as mean square error loss between clean and enhanced magnitude features.

2.3.2 Short Term Objective Intelligibility

Short Term Objective Intelligibility(STOI) measures the intelligibility of the enhanced audio within the range [0, 1] with high values corresponding to strong intelligibility. The method computes DFT for both clean and enhanced audio and group them into 15 one-third octave bands. Each band is called an Time-Frequency unit; an intermediate intelligibility measure of each unit depends on N consecutive units. The global intelligibility is the average of all intermediate measures. (Alonso et al., 2021) shows that STOI is quite accurate to estimate people's perception. We use STOI score to evaluate the performance on dev set. The model with best STOI on dev set are picked to enhance test audios.

2.3.3 Negative Log-Likelihood Loss

The pretrained ASR model is seq2seq based. To fine-tune the model with and without speech enhancement module, we use NLLLoss for both cases.

3 Experiments & Results

3.1 Training Setup

All the experiments were run on a single GPU. For both speech enhancement training, and ASR fine-tuning, we choose Adam optimizer with $1e^{-4}$ learning rate, which converges faster than SGD. It takes 5 hours to train 100 epochs of the SE module, and 7.5 hours each to fine-tune 10 epochs of ASR model with and without SE module.

Validation takes the most time during fine-tuning due to our beam-search relies on a Attention RNN language model that comes with the pretrained ASR model. Since the ASR model is separately fine-tuned, we generate enhanced audio using our SE module in advance, and simply load them in training.

3.2 Dataset

We use LibriMix as our dataset in both training and evaluation process. LibriMix is an open source dataset that is derived from the combination of LibriSpeech signal (clear signal) and WHAM noise. LibriMix can be generated for several speakers, with one or more layers of noise added. In this study, we generate single source LibriMix, which means that the audio is a combination of one layer of noise and one speaker speaking. There are 13000 training samples, 3000 dev samples and 3000 test samples; the training data consists of 54 hours of audio with an average SNR of 3.1.
LibriMix itself does not provide the ground truth transcript for an audio. However, since LibriMix is derived from LibriSpeech that offers the ground truth transcript for each audio, we extracted the transcript from LibriSpeech to be the ground truth of the corresponding LibriMix audio.

3.3 Baseline

Table 1 list the performance of the pretrained ASR model(Ravanelli et al., 2021) on our clean and noisy audio set. The model performs well in clean condition, but significantly worse in noisy conditions. On the other hand, it shows a huge room to improve.

Table 1: Pre-trained ASR model WER

signal	WER
clean audio	3.96
noisy audio	42.23

3.4 Speech Enhancement Result

Table 2: WER using speech enhancement module over time

Dev Set STOI	WER
0.75	33.43
0.79	28.57
0.82	24.19

The model with highest STOI score on dev set is picked for enhancing the test audio. Table 2 shows the performance with the same SE module under different number of epochs. The WER improvement in terms of STOI is linear, suggesting that while our SE module reaches its capacity limit, it is still under-powered to fully remove the noise. Due to our computation constraint, we do not further investigate stronger SE architecture; however, even with under-powered model, there has been a 42% error reduction from the baseline. We observe that enhanced audios have higher speech volume standing out from the background noise, compared to noisy audio where speech volume is roughly the same as the noise.

3.5 Fine Tuning Result

Table 3: WER after fine-tuning

With/Without SE	WER
With SE	18.83
Without SE	24.34

Table 3 shows the fine-tuning performance with and without the use of SE module. By fine-tuning the ASR model, there has been a 22% improvement compared to only using SE module and a 55% improvement over baseline. It is not surprising to fine-tuning leads to better performance as SE module tends to introduce its own characteristics to the enhanced audio which could be captured through fine-tuning.

On the other hand, fine-tuning the model directly on noisy data also yields significant improvement. The performance is similar to including a SE module, which could suggest that ASR model is learning to mask the noise. This opens up a direction to study the transfer learning of ASR on noisy dataset without extra modules. Yet including a SE module still has the huge benefit of easy-plugin and faster training. ASR model is in general more complex and takes longer to train than SE module. In our experiment, it takes double time to fine-tune an ASR model than to train a SE module from scratch, and ends up with similar performance. One could use a pretrianed SE module to avoid any further computation at all.

4 Conclusion

Our studies show that ASR model trained on clean speech suffer significant loss from background noises. Introducing a separate speech enhancement as the front end of ASR is effective in reducing the negative impact, and fine-tune the ASR model using enhanced audio yields best result. While fine-tuning the ASR model on noisy data could potentially match the performance of using a speech enhancement module, the latter still has the advantages of easy-to-use and fast-to-train.

References

A. Alonso, V. García, I. Hernaez, E. Navas, and J. Sanchez. Automatic Speaker Adaptation Assessment Based on Objective Measures for Voice Banking Donors. In *Proc. IberSPEECH 2021*, pages 210–214, 2021. doi: 10.21437/IberSPEECH.2021-45. URL `http://dx.doi.org/10.21437/IberSPEECH.2021-45`.

Z. Chen, Y. Huang, J. Li, and Y. Gong. Improving mask learning based speech enhancement system with restoration layers and residual connection. In *Interspeech*. ISCA, August 2017. URL `https://www.microsoft.com/en-us/research/publication/improving-mask-learning-based-speech-enhancement-system-restoration-layers-residual-connection`

J. Cosentino, M. Pariente, S. Cornell, A. Deleforge, and E. Vincent. Librimix: An open-source dataset for generalizable speech separation, 2020.

A. Narayanan and D. Wang. Investigation of speech separation as a front-end for noise robust speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(4):826–835, 2014. doi: 10.1109/TASLP.2014.2305833.

M. Ravanelli, T. Parcollet, A. Rouhe, P. Plantinga, E. Rastorgueva, L. Lugosch, N. Dawalatabad, C. Ju-Chieh, A. Heba, F. Grondin, W. Aris, C.-F. Liao, S. Cornell, S.-L. Yeh, H. Na, Y. Gao, S.-W. Fu, C. Subakan, R. De Mori, and Y. Bengio. Speechbrain. urlhttps://github.com/speechbrain/speechbrain, 2021.

M. Strake, B. Defraene, K. Fluyt, W. Tirry, and T. Fingscheidt. Speech enhancement by lstm-based noise suppression followed by cnn-based speech restoration. *EURASIP Journal on Advances in Signal Processing*, 2020(1):49, Dec 2020. ISSN 1687-6180. doi: 10.1186/s13634-020-00707-1. URL `https://doi.org/10.1186/s13634-020-00707-1`.