Reinforcement Learning-Enhanced ASR Systems: A Novel Reward-Based Framework

Anonymous Author(s)
Affiliation
Address
email

Abstract

Automatic Speech Recognition (ASR) systems often face challenges in handling diverse and noisy speech data. To address this, we investigate how data augmentation techniques can enhance ASR performance. This study applies reinforcement learning (RL) methods to optimize the data augmentation process, aiming to improve recognition accuracy. Specifically, we employ a Deep Q-Network (DQN) to dynamically select the best augmentation strategies during training, considering both static properties of the audio waveform and extracted features. The methodology involves defining states based on audio and feature properties, actions as different augmentation techniques, and rewards derived from improvements in training loss, Character Error Rate (CER), and Signal-to-Noise Ratio (SNR). We incorporate instant rewards to fine-tune the feature extraction process, ensuring immediate adjustments based on variance and SNR differences. Our approach includes a reward distribution mechanism that balances instant rewards with cumulative improvements in CER and Word Error Rate (WER), ensuring robust performance enhancements. Experimental results demonstrate significant enhancements in ASR accuracy and robustness.

1 Introduction

Automatic Speech Recognition (ASR) systems have significantly advanced in recent years, driven by deep learning techniques [1]. However, handling diverse and noisy speech data remains a challenge. Recent studies have explored various data augmentation methods to address this issue. Over the years, primarily transitioning from traditional Hidden Markov Model - Gaussian Mixture Model (HMM-GMM) based systems to modern deep learning-based systems.

Learning-based systems leverage neural networks to achieve superior accuracy and robustness. These systems typically employ models such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Long Short-Term Memory networks (LSTMs), and Transformer models. One of the well-developed learning-based ASR systems is SpeechBrain, an open-source PyTorch toolkit developed by a community of professionals. In industry, SpeechBrain has been utilized in various industry projects and products by companies such as Facebook and Hugging Face to enhance ASR capabilities and facilitate research. Its versatility and modularity framework have made it a popular choice among researchers for validating ASR algorithms, as evidenced by its citation in over 100 research papers since its release.An overview of the learning-based ASR system workflow is depicted in Figure 1.

The process starts with feature extraction from the raw audio, followed by data augmentation and feature augmentation. The features are then encoded and embedded into a higher-dimensional space. The embedded features are decoded using beam search to generate the final text predic-

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.

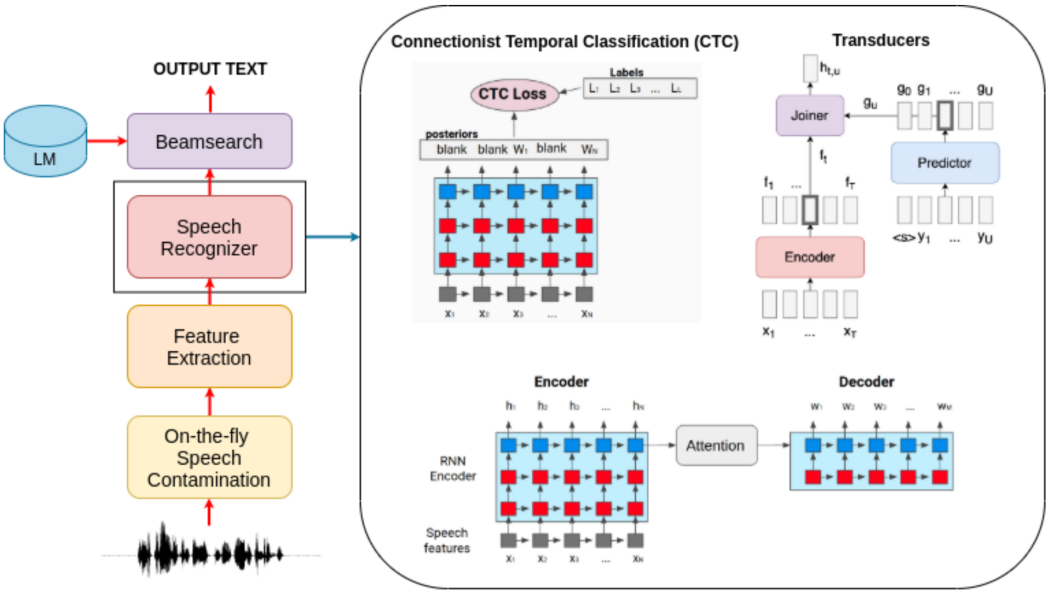

Figure 1: Learning-based ASR workflow

tion. Throughout the training process, objectives such as training loss and evaluation metrics like Character Error Rate (CER) and Word Error Rate (WER) are used to optimize the model.

Numerous approaches have been proposed to address the noise issue in ASR systems. One category of these approaches is speech enhancement (SE), which aims to generate enhanced speech signals that closely match clean and undistorted speech signals by removing the noise components from noisy speech. Traditional SE methods are designed based on certain assumptions about speech and noise characteristics, which may yield satisfactory performance in terms of speech quality but do not directly improve the ASR performance.

Numerous approaches have been proposed to address the noise issue in ASR systems. One category of these approaches is speech enhancement (SE), which aims to generate enhanced speech signals that closely match clean and undistorted speech signals by removing the noise components from noisy speech. Traditional SE methods are designed based on certain assumptions about speech and noise characteristics, which may yield satisfactory performance in terms of speech quality but do not directly improve the ASR performance.[2]

Recently, deep learning-based SE approaches have received increased attention due to their ability to effectively transform noisy speech into clean speech or to accurately estimate masks to filter out noise components.[3] These models typically use mean square error (MSE) as the objective function for training. However, although MSE-based objectives are effective for noise reduction, they do not necessarily improve speech quality, intelligibility, or ASR performance.[4]

In light of these limitations, it is evident that recognition results should be used as the optimal objective function for SE when the goal is to achieve good ASR performance. However, this is challenging due to the complexity and non-differentiability of ASR systems, which consist of multiple modules such as acoustic and language models. Additionally, building a robust ASR system requires significant resources, making it beneficial to use a well-established ASR system from a third party.

In this study, we propose to adopt the reinforcement learning (RL) algorithm to optimize SE models based on recognition results. The main concept of RL is to take actions in an environment to maximize a cumulative reward. Unlike supervised and unsupervised learning algorithms, RL algorithms learn to achieve complex goals iteratively, and evaluated the proposed RL-based SE system on a LibriSpeech ASR corpus dataset, which is widely recognized for its reliability and comprehensive coverage of spoken English.

Specifically, we implement a Deep Q-Network (DQN) to choose appropriate wave augmentation and feature augmentation actions, optimizing the SE model. The reward structure includes improvements in character error rate (CER) and word error rate (WER), as well as instance-based rewards.

2 Related Works

In the field of speech signal recognition, data augmentation techniques play a critical role in enhancing the robustness of speech recognition systems. These techniques can be categorized into *augmentation techniques* and *feature extraction methods*. Additionally, reinforcement learning (RL) has been successfully applied to speech enhancement tasks.

2.1 Augmentation Techniques

Augmentation techniques apply transformations to both the raw audio signals and extracted speech features to generate more varied training data. The mainly techniques are add noise, shifting, warping,speech perturbation and codec augmentation.

Codec augmentation simulates the effects of different audio codecs to increase data diversity. This technique involves applying random audio codecs to input waveforms, thereby enhancing the robustness of the model by exposing it to various types of audio distortions.

Speed perturbation changes the playback speed of the audio signal to simulate different speaking rates. This is achieved by using different sampling rates during the resampling process, thereby altering the duration and frequency characteristics of the audio signal without changing its pitch.

Warp applies random warping to the time or frequency axis, which can be achieved by:

$$\mathbf{X}_{\text{warp}}(t, f) = \mathbf{X}(g(t), f)$$

where $g(t)$ is a non-linear function that warps the time axis.

2.2 MFCC Feature Extraction

Mel-Frequency Cepstral Coefficients (MFCCs) are widely used in speech recognition for their ability to represent the short term power spectrum of sound. The MFCC feature extraction process involves several steps, pre-emphasis, framing, windowing, discrete fourier transform (DFT), mel-scale fitering, logarithm, discrete cosine Transform (DCT). Mathematically, the MFCCs can be represented as:

$$\text{MFCC}(n) = \sum_{m=1}^{M} \log(E_m) \cos\left[\frac{\pi n(m-0.5)}{M}\right]$$

where E_m is the energy of the m-th filter bank and M is the total number of filter banks.

2.3 Deep Q-Network (DQN) for Speech Enhancement

Reinforcement learning (RL) has been applied to speech enhancement tasks, where it learns action strategies to maximize the cumulative reward related to speech recognition performance. The Deep Q-Network (DQN) is particularly effective in this context due to its capability to handle discrete actions and its ease of training.

In DQN, the Q-value function $Q(s, a)$ represents the expected cumulative reward for taking action a in state s. The update rule for the Q-value function is:

$$Q(s, a) \leftarrow Q(s, a) + \alpha \left[r + \gamma \max_{a'} Q(s', a') - Q(s, a)\right]$$

where α is the learning rate, γ is the discount factor, r is the immediate reward, and s' is the new state after taking action a.

Assuming an augmentation operation set $\mathcal{A}$, DQN selects an optimal augmentation operation a^* to apply to the current audio features to maximize future cumulative rewards:

$$a^* = \arg\max_{a \in \mathcal{A}} Q(s, a)$$

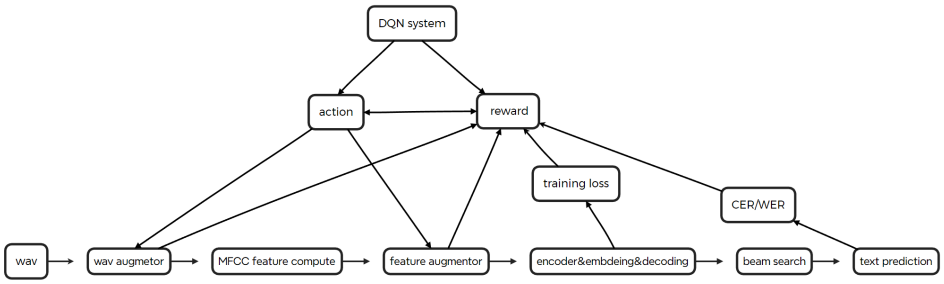

Figure 2: The block diagram of the proposed SE system.

3 Proposed Method

3.1 Training framework

The training framework, shown in Figure 2, leverages reinforcement learning (RL) to optimize the speech enhancement (SE) model. The RL agent uses feedback from recognition results, based on metrics like Character Error Rate (CER) and Word Error Rate (WER), to iteratively improve the SE model's performance.

In the core of the framework lies the reinforcement learning module, which consists of the Deep Q-Network (DQN) system. The DQN system is responsible for determining the optimal sequence of actions (audio processing steps) to enhance the input waveform. The actions chosen by the DQN system modify the features or directly the waveform itself, depending on the current state of the audio signal.

3.2 Action estimation module

In this reinforcement learning framework, the action setting is divided into two main parts: wave augmentation and feature augmentation.

Wave augmentation. Includes 10 different wave operations, including adding noise, time stretching, volume adjustment.The selection of these operations is based on the current wave statistical properties, which describe the characteristics of the audio signal.

Feature Augmentation. Includes 6 different feature augmentation operations, including shifing, drop and warping. The selection of these operations is based on the current feature properties, MFCC calculation, spectrum filter result or gabot convolution. Mathematically, the action space can be described as:

$$A = \{a_1, a_2, \ldots, a_{10}\} \cup \{a_{11}, a_{12}, \ldots, a_{16}\}$$

where the first 10 actions correspond to wav augmentation operations, and the last 6 actions correspond to feature augmentation operations.

3.3 Reward estimation module

The reward design includes the instant reward,and total reward. Instant reward based on immediate feedback after wavs and features implement augmentation,calculation by signal SNR and variance difference. Total rewards can also be calculated based on improvements in training loss and error. Specifically, rewards can be given based on improvements in CER (Character Error Rate) or WER (Word Error Rate) and training loss.

Training loss. Composed of CTC probabilities, generated by the Connectionist Temporal Classification (CTC) layer,helping the model predict the most likely sequence of outputs based on the input, and sequence probabilities, the decoder generates the output sequence step-by-step, which represent the likelihood of each possible next token at each step in the output sequence, guiding the generation of the final recognized text.

The reward function $R(s, a)$ can be defined as follows:

$$R(s, a) = -(\Delta\text{CER} + \Delta\text{WER} + \Delta\text{Training Loss}) + \text{SNR Difference}$$

4 Experiment

The main modification of these modules is in `templets->ASR->asr_training, DQN_agent`.The datasets used for the experiments are from Librispeech. The dimensions of the DQN network layers are as follows: Layer 1 has 64 inputs and 64 outputs, Layer 2 has 32 inputs and 32 outputs, and Layer 3 has 32 inputs and outputs an action dimension.The action dimension is set to 10 in the wave augmentor and 6 in the action dimension.

4.1 Results

The performance comparison between the baseline model and the RL-augmentor model is as follows: The baseline model has a training loss of 1.3, a Character Error Rate (CER) of 5.36%, and a Word Error Rate (WER) of 15.24%. The RL-augmentor model has a training loss of 0.96,a Character Error Rate (CER) of 4.18%, and a Word Error Rate (WER) of 14.14%.

5 Conclusion

In this study, we proposed a reinforcement learning (RL)-based speech enhancement (SE) system aimed at improving automatic speech recognition (ASR) performance in noisy environments. By employing a Deep Q-Network (DQN) to dynamically select optimal data augmentation strategies during training, our approach effectively reduces Character Error Rate (CER) and Word Error Rate (WER). Experimental results demonstrate that the RL-based SE system achieves a training loss reduction from 1.3 to 0.96 and a significant reduction in CER by 5.36% and WER by 15.24%, without requiring retraining of the ASR system.

6 Discussion

Despite the promising results, this approach effectively addresses the limitations of traditional SE methods that rely on fixed assumptions about speech and noise characteristics. Future work could explore the impact of more complex reward structures and test the RL-based SE system in diverse real-world noisy environments and with different languages. Additionally, incorporating more strategies for embedding and decoding choices might provide richer information for decision-making and potentially improve the robustness of the ASR system.

References

[1] Hao Wang et al. "Unifying Robustness and Fidelity: A Comprehensive Study of Pretrained Generative Methods for Speech Enhancement in Adverse Conditions". In: *arXiv preprint arXiv:2309.09028* (2023).

[2] Anurag Pandey, DeLiang L. Wang, and Yuxuan Yang. "Time-Domain Speech Enhancement for Robust Automatic Speech Recognition". In: *arXiv preprint arXiv:2305.18524* (2023).

[3] SyuSiang Wang et al. "Reinforcement Learning Based Speech Enhancement for Robust Speech Recognition". In: *arXiv preprint arXiv:1811.04224* (2018). URL: `https://arxiv.org/abs/1811.04224`.

[4] Yun-Hsuan Lai et al. "Deep Learning-Based Speech Enhancement With a Loss Trading Off the Speech Distortion and the Noise Residue for Cochlear Implants". In: *Frontiers in Medicine* 8 (2021). URL: `https://www.frontiersin.org/articles/10.3389/fmed.2021.740123/full`.