Adversarial Latent Representation Learning for Speech Enhancement

Yuanhang Qiu, Ruili Wang*

School of Natural and Computational Sciences, Massey University, New Zealand
{y.qiu1,ruili.wang}@massey.ac.nz

Abstract

This paper proposes a novel adversarial latent representation learning (ALRL) method for speech enhancement. Based on adversarial feature learning, ALRL employs an extra encoder to learn an inverse mapping from the generated data distribution to the latent space. The encoder builds an inner connection with the generator, and provides relevant latent information for adversarial feature modelling. A new loss function is proposed to implement the encoder mapping simultaneously. In addition, the multi-head self-attention is also applied to the encoder for learning of long-range dependencies and further effective adversarial representations. The experimental results demonstrate that ALRL outperforms current GAN-based speech enhancement methods.

Index Terms: adversarial feature learning, latent space, speech enhancement

1. Introduction

Speech enhancement aims to improve the intelligibility and overall perceptual quality of contaminated speech signals [1]. There are many practical applications such as telephone communications [2], hearing-aid devices [3], and human-computer interactions [4], which regard the speech enhancement as an essential operation for different purposes and processing stages. Obviously, more complicated and critical application scenarios require higher performance of speech enhancement.

Classic signal processing methods of speech enhancement (e.g. Wiener filtering [5], spectral subtraction [6]) perform well in specific additive noise suppressing. However, these methods are difficult to process assorted unknown noise interference satisfactorily. In order to solve this problem, learning appropriate representation of noise data distribution is a key procedure in current data-driven approaches.

Recently, deep learning based methods have shown revolutionary information learning and reconstruction property in many research areas. Profiting from this, a series of neural network based speech enhancement methods such as denoising autoencoder [7, 8], LSTM-based [9], CNN-based methods [10, 11] were also developed for improving speech enhancement performance. Particularly, the generative adversarial networks (GAN) [12, 13], which was originally proposed with artful architecture design for high quality images generation in computer vision, has been applied successfully to speech enhancement [14, 15].

GAN consists of a generator and a discriminator, which are trained adversarially up to the Nash Equilibrium [12]. For speech enhancement as shown in Figure 1, the generator usually takes in noisy speech and extra noise distributions (i.e. latent vectors) as input and exports targeted data distribution. The discriminator is considered as a classifier trained to distinguish generated samples and clean speech as fake or true. With effective representation learning and enhancement performance, GAN-based speech enhancement methods have attracted a good proportion of attention[14, 16] in speech enhancement.

However, no attention has been paid to latent space for representation learning in speech enhancement. Initially, GAN can generate high quality target from latent vectors based on real data distribution. Thus, in speech enhancement, our hypothesis is that the latent vectors play an important part for representation learning conditioned on explicit noisy data distribution.

In this work, we propose a novel GAN-based method named adversarial latent representation learning (ALRL), which employs an extra encoder inversely mapping the generated data distribution to the latent space, for speech enhancement improvement. In particular, the encoder attempts to build an inner connection with the generator and provides relevant representation information for the adversarial features modelling. The new architecture remodels the inner projection from the concatenated input of noisy speech and latent vectors to the clean speech distribution. To implement the encoder mapping, we propose a new encoder mapping loss function, which captures latent representation by calculating the squared Euclidean distance from the inverse mapped generator samples to the latent vectors. Also, we combine the encoder loss with the relativistic loss [17] to further improve the effectiveness of information learning between the generator and discriminator. In the meanwhile, the multi-head self-attention mechanism [18] is also applied to the encoder in our ALRL for long-range dependencies capturing and further effective representation learning.

The remainder of this paper is organized as follows. The related work is given in Section 2. The details of our adversarial latent representation learning are introduced in Section 3. Sections 4 gives the design of the experiments, while the experimental results are presented in Section 5. Finally, the conclusions and future work are shown in Section 6.

2. Related Work

In this section, we introduce related GAN-based speech enhancement methods and present a preliminary investigation of latent space with GANs. Based on these works, our ALRL is proposed to learn semantic representation and improve speech enhancement performance.

2.1. GAN-based Speech Enhancement

Recently, the GAN-based models have derived huge progress on semantic representation learning and improved speech enhancement performance significantly. Speech Enhancement GAN (SEGAN) is one of the most famous frameworks proposed for time-domain speech enhancement with improved conditional GAN [14], which combined the conditional GAN with the least-squares GAN (LSGAN) together to further alleviate vanishing gradients. This modification is proved to be ef-
effective in performance improvement. Below is the loss function of its discriminator:

\[
L_D = \frac{1}{2} E_{x \sim P_{\text{data}}, z \sim P_z} [(D(x, x_e) - 1)^2] + \frac{1}{2} E_{x \sim P_{\text{data}}, x_e \sim P_{\text{latent}}} [(D(G(z, x_e), x_e)) - 1]^2
\]
and its generator:

\[
L_G = \frac{1}{2} E_{x \sim P_{\text{data}}, x_e \sim P_{\text{latent}}} [(D(G(z, x_e), x_e) - 1)^2]
\]

where \( x_e \) denotes noisy speech; \( x \) denotes clean speech; and \( z \) denotes random noise distribution (i.e. latent information).

SEGAN operated on raw speech waveform directly rather than the processed spectral features, which is considered to be able to preserve original sequential information such as phase information effectively. SEGAN worked end-to-end and was trained adversarially based on GAN. The fully convolutional architecture consists of a downsampling and a upsampling modules (i.e. encoder and decoder). The random noise \( z \) (i.e. latent vectors) was added to the bottleneck layer for information compensation whereas without further introduction of it. As we know, this work applied conditional GAN to speech enhancement firstly and obtained outstanding performance. In the meanwhile, conventional GANs were also applied to noise-robust speaker verification [19] and speech recognition [20]. However, the latent space still has not been explored thoroughly in speech signal processing.

Speech Enhancement Relativistic GAN (SERGAN) [16] is another framework exploring speech enhancement based on GAN. In the standard GAN, the discriminator is developed to estimate the probability that the original data is real and the generated data is fake, on the contrary, the generator is trained to increase the probability that fake data is real. However, it should simultaneously decrease the probability that real data is real when the generator learns to increase that probability. To accomplish the assumption, the relativistic GAN [17] was proposed for more stable model and higher quality data samples. Below is the loss function of discriminator:

\[
L_D = -E_{x \sim P_{\text{data}}, x_e \sim P_{\text{latent}}} [\log(\sigma(C(x, x_e) - C(G(z, x_e), x_e)))]
\]
and generator:

\[
L_G = -E_{x \sim P_{\text{data}}, x_e \sim P_{\text{latent}}} [\log(\sigma(C(G(z, x_e), x_e) - C(x, x_e)))]
\]

where \( \sigma \) is the sigmoid non-linearity, and \( C(x) \) denotes the discriminator without the final sigmoid layer. \( D(x) = \sigma(C(x)) \).

SERGAN built closer information connection between the generator and discriminator for speech enhancement. Moreover, the gradient penalty was also utilized to stabilize model training and improve enhancement performance. The method held a similar architecture with SEGAN but adopted a new loss function to boost information communication between the generator and discriminator. However, this work still did not explore latent space further.

In addition, the improved Speech Enhancement GAN (iSEGAN) [21] conducted a preliminary experiment to explore the impact of the latent vectors on a speech enhancement model by comparing the performance of the model trained with and without latent vector. The results showed that the latent vectors could slightly affect the model performance but were helpful to stabilize model training.

These methods mentioned above attempt to obtain performance gains by modifying model architecture. Also, there are works that explore data space for better model performance.

**2.2. Latent Space**

It is possible that the latent vectors are used by the generator in a highly entangled way, causing the individual dimensions of latent vectors to not correspond to semantic features of the data [22]. For image generation, Chen et al. proposed to adopt a mutual information strategy for inducing latent vectors. The method decomposed the input noise vectors into a set of semantically meaningful factors of variation rather than using single unstructured noise vectors. The work discovered that these latent factors can target salient semantic features of data distribution.

Similarly, Donahue et al. noticed that GAN models could capture semantic variation from latent space, however, have no means of projecting data back into the latent space [23]. This resulted in the architecture ignoring much of the useful information presenting in the structure of the data itself. In addition, interpolations in the latent space of the generator produced smooth and plausible semantic variations and made the model learns associating particular latent directions with specific features. Thus, the Bidirectional Generative Adversarial Networks (BiGAN) was proposed to learn generative mapping from simple latent distributions to arbitrarily complex data distribution [23]. Another similar work about latent space exploration was proposed to map training examples in the data space to the space of latent variables as well [24].

Inspired by mentioned work, we infer that the latent space plays an important role in generative model for semantic representation capturing. Thus, we propose adversarial latent representation learning method for speech enhancement.

**3. Adversarial Latent Representation Learning**

In this section, we introduce our adversarial latent representation learning (ALRL) for speech enhancement.

**3.1. ALRL**

The related works show that the latent space can target salient semantic features of data distribution and provide effective guidance for information learning. In our work, an encoder is built for latent representation learning whereas our encoder will be trained for inverse mapping from generated samples to latent space. To implement the encoder mapping, we propose a new encoder loss function, which captures latent representation
by calculating the squared Euclidean distance from the inverse mapped generator samples to the latent vectors.

To further improve the effectiveness of information learning, we combine the encoder loss with the relativistic loss function [17]. The encoder, generator and discriminator will be trained simultaneously. Below is our new loss function for the generator:

\[ L_G = - E_{x \sim P_x, \epsilon \sim P_{\epsilon}} [\log(C(G(z, x), x) - C(x, x))] - E_{x \sim P_x} ||E(G(z, x)) - z||^2_2 \]  

(5)

where \( E \) is defined by calculating the squared Euclidean [25] loss. The new loss function improves the semantic representation learning of the generator. To avoid vanishing gradients further, the gradient penalty regularization is also used in discriminator as proposed in [16]. Below is the discriminator:

\[ L_D = - E_{x \sim P_x, \epsilon \sim P_{\epsilon}} [\log(C(x, x) - C(G(z, x), x))] - \lambda E_{\tilde{x} \sim P_{\tilde{x}}, \epsilon \sim P_{\epsilon}} [||\nabla_{\tilde{x}, \epsilon} C(\tilde{x}, x)||_2^2 - 1]^2 \]  

(6)

where \( P(\tilde{x}, \epsilon) \) is the joint probability of \( \tilde{x} = \epsilon x + (1 - \epsilon)G(z, x) \) and \( x; \epsilon \) is sampled from a uniform distribution in [0, 1]; \( \lambda \) is the hyper-parameter that controls the gradient penalty.

Similar setup as SEGAN, the generator receives the noisy speech signal and latent vectors and put them into multi-layer convolutions with the filter (width = 31 and strides N = 2). Before the intermediate layer, a normal 2D convolutional followed by parametric rectified linear units (PReLU) [26] is used for inherent information capturing from input distributions. Then 2D transposed convolutional, followed again by PReLU, is used for information reconstruction.

The discriminator is considered as a binary classifier for judgement to real samples and generated samples. The main component is 2D convolutional layer as well. Differently, the discriminator applies the LeakyReLU function [27] and virtual batch normalization function rather other only PReLUs in the generator. This will greatly improve the discriminable information learning of discriminator and alleviate gradients vanishing.

The main structural ingredients of the encoder are also the 2D convolutional layer. Especially, the multi-head self-attention layer is applied to the encoder for the specific speech information locating and the long-range dependencies learning [18].

3.2. Multi-head Self-attention

For each input sequence, the (Query, Key, and Value) vectors will be created by applying learned linear projection or using feed-forward layers. Then the attention will be applied to all other positions with the three vectors. The procedure can be described as below:

\[ \text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}}) \cdot V \]  

(7)

where \( d_k \) is the dimension of the key vectors. The purpose of this scaling is to improve numerical stability as the dimensions of keys, values, and queries grow. The obtained attention at each position will be used to times the value vector of all other positions including itself. This will produce multiple results called multi-head attention. The sum of all heads will be the final result of the first position input. The same operation will be applied at each subsequent position. Below is the equation of the multi-head calculation:

\[ \text{MultiHead}(Q, K, V) = (\sum_{i=1}^{h} \text{head}_i)W^O \]  

(8)

4. Experiments

4.1. Dataset

The selected database is an open and standard resource for the performance evaluation of a speech enhancement system. The original clean speech was selected from Voice Bank corpus, including 28 speakers – 14 males and 14 females with the same accent region (England). There are two artificially generated noises (i.e. speech-shape noise and babble) and eight real noises extracted from the Diverse Environments Multi-channel Acoustic Noise Database (DEMAND) database.

For training data, the signal-to-noise (SNR) values are 15dB, 10dB, 5dB and 0dB. That signifies 40 different noisy conditions are produced in this corpus. Each speaker contributes 10 sentences, the corpus will add 400 sentences in total. Each clean speech waveform needs to be normalized and trimmed off silence segments of which are longer than 200ms at the beginning and the end.

Another two speakers (a male and a female), not including in the training data, are picked as the test data from the Voice Bank corpus with the same England accent. Five other noisy types were selected from the DEMAND database. The SNR values are 17.5dB, 12.5dB, 7.5dB and 2.5dB, respectively. Thereby, there are 20 different conditions for each sentence of per test speaker.

4.2. Setup

ALRL adopts the Adam optimizer [29], a learning rate of 0.0002. The raw speech waveforms preserving the original inherent content of speech signals are used same as SEGAN [14]. About one second speech chunks (16384 samples) is segmented by a sliding window (500ms overlap) during training, however no overlap during test. In addition, a high-frequency

\[ \text{https://datashare.is.ed.ac.uk/} \]
Table 1: The evaluation results of different methods in quality and intelligibility. The results include four SNR conditions (i.e. 17.5 dB, 12.5 dB, 7.5 dB, and 2.5 dB) and the overall values. The best results obtained are highlighted in the bold font.

<table>
<thead>
<tr>
<th>Strategies</th>
<th>Methods</th>
<th>SNR</th>
<th>PESQ</th>
<th>CSIG</th>
<th>CBAK</th>
<th>CVOL</th>
<th>CSII$_{\text{high}}$</th>
<th>CSII$_{\text{mid}}$</th>
<th>NCM</th>
<th>STOI</th>
</tr>
</thead>
<tbody>
<tr>
<td>SEGAN [14]</td>
<td>17.5dB</td>
<td>2.60</td>
<td>4.88</td>
<td>3.28</td>
<td>3.26</td>
<td>99.67</td>
<td>95.60</td>
<td>99.39</td>
<td>95.43</td>
<td></td>
</tr>
<tr>
<td></td>
<td>12.5dB</td>
<td>2.29</td>
<td>4.76</td>
<td>3.06</td>
<td>2.96</td>
<td>99.12</td>
<td>91.10</td>
<td>98.91</td>
<td>94.25</td>
<td></td>
</tr>
<tr>
<td></td>
<td>7.5dB</td>
<td>2.06</td>
<td>4.51</td>
<td>2.87</td>
<td>2.69</td>
<td>97.68</td>
<td>85.21</td>
<td>97.18</td>
<td>92.91</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2.5dB</td>
<td>1.76</td>
<td>4.03</td>
<td>2.59</td>
<td>2.35</td>
<td>93.10</td>
<td>74.78</td>
<td>92.81</td>
<td>89.03</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Overall</td>
<td>2.16</td>
<td>3.48</td>
<td>2.94</td>
<td>2.80</td>
<td>97.29</td>
<td>86.37</td>
<td>96.98</td>
<td>92.80</td>
<td></td>
</tr>
<tr>
<td>SERGAN [16]</td>
<td>17.5dB</td>
<td>2.95</td>
<td>4.95</td>
<td>3.56</td>
<td>3.51</td>
<td>99.75</td>
<td>96.93</td>
<td>99.69</td>
<td>96.11</td>
<td></td>
</tr>
<tr>
<td></td>
<td>12.5dB</td>
<td>2.67</td>
<td>4.92</td>
<td>3.31</td>
<td>3.21</td>
<td>99.36</td>
<td>93.36</td>
<td>99.41</td>
<td>95.70</td>
<td></td>
</tr>
<tr>
<td></td>
<td>7.5dB</td>
<td>2.43</td>
<td>4.77</td>
<td>3.09</td>
<td>2.97</td>
<td>98.21</td>
<td>87.89</td>
<td>98.49</td>
<td>93.73</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2.5dB</td>
<td>2.09</td>
<td>4.39</td>
<td>2.79</td>
<td>2.61</td>
<td>94.63</td>
<td>78.38</td>
<td>95.91</td>
<td>90.19</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Overall</td>
<td>2.52</td>
<td>4.75</td>
<td>3.18</td>
<td>3.06</td>
<td>97.91</td>
<td>88.89</td>
<td>98.33</td>
<td>93.69</td>
<td></td>
</tr>
<tr>
<td>Our ALRL</td>
<td>17.5dB</td>
<td>3.00</td>
<td>4.97</td>
<td>3.65</td>
<td>3.60</td>
<td>99.76</td>
<td>96.98</td>
<td>99.71</td>
<td>96.12</td>
<td></td>
</tr>
<tr>
<td></td>
<td>12.5dB</td>
<td>2.73</td>
<td>4.94</td>
<td>3.36</td>
<td>3.32</td>
<td>99.39</td>
<td>93.50</td>
<td>99.40</td>
<td>95.11</td>
<td></td>
</tr>
<tr>
<td></td>
<td>7.5dB</td>
<td>2.47</td>
<td>4.81</td>
<td>3.13</td>
<td>3.06</td>
<td>98.28</td>
<td>88.03</td>
<td>98.61</td>
<td>93.67</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2.5dB</td>
<td>2.14</td>
<td>4.45</td>
<td>2.82</td>
<td>2.70</td>
<td>94.81</td>
<td>78.80</td>
<td>96.20</td>
<td>90.31</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Overall</td>
<td>2.57</td>
<td>4.79</td>
<td>3.23</td>
<td>3.16</td>
<td>97.98</td>
<td>89.08</td>
<td>98.43</td>
<td>93.71</td>
<td></td>
</tr>
</tbody>
</table>

5. Results

Many objective evaluation measures can evaluate enhanced speech performance with high correlation. The Perceptual Evaluation of Speech Quality (PESQ: from -0.5 to 4.5) for wideband speech is an effective full-reference speech quality evaluation algorithm [30]. Moreover, we also implement the composite evaluation metrics of the enhanced speech including the predicted Mean Opinion Score (MOS) of signal distortion (CSIG: from 1 to 5), background noise distortion (CBAK: from 1 to 5), and overall quality (CVOL: from 1 to 5).

The intelligibility of enhanced speech is also implemented in this work. The Coherence-based Speech Intelligibility Index (CSII) measure is computed for the medium-level (CSII$_{\text{mid}}$) and high-level (CSII$_{\text{high}}$) segments of each speech sentence, which can predict the intelligibility of peak-clipping and centering-clipping distortion in the speech signal [31]. In addition, another popular speech intelligibility evaluation metrics the Normalized Covariance Metric (NCM) [31] and the Short-Time Objective Intelligibility (STOI) [32] are also conducted.

The experimental results of different methods are shown in Table 1. We set the SEGAN method as the baseline and its result was described in paper [14]. According to the description of the SERGAN method [16], we retrain the SERGAN model and obtained the results as shown in Table 1. Besides the overall evaluation results, we also split the test data to four respective SNR conditions (i.e. 17.5dB, 12.5dB, 7.5dB, and 2.5dB) and obtain evaluation results.

As shown in Table 1, our adversarial latent representation learning (ALRL) method outperforms the SEGAN and SERGAN methods and achieves the highest scores in both speech quality and intelligibility. Specifically, our method improves PESQ by 1.98% and 19.0%, improves STOI by 0.213‰ and 9.81‰ over SERGAN and SEGAN, respectively. Moreover, our method also obtains outstanding enhancement performance in each SNR condition. Our method improves PESQ by 1.69% and 15.4% in 17.5dB, 2.39% and 21.6% in 2.5dB, improves STOI by -0.208‰ and 7.23‰ in 17.5dB, 1.33‰ and 14.4‰ in 2.5dB over SERGAN and SEGAN. Our ALRL can effectively improve the intelligibility and quality of noisy speech, especially for low SNR scenarios.

6. Conclusions and Future Work

In this paper, we propose a novel adversarial latent representation learning (ALRL) method for speech enhancement. An extra encoder model is built in our ALRL to learn the semantic representation by inverse mapping from the generated samples to the latent space. The encoder greatly improves the effectiveness of adversarial training and the complex data distribution learning. To accomplish the inverse mapping, we propose a new loss function, which captures latent representation by calculating the squared Euclidean distance from the inverse mapped generator samples to the latent vectors. In addition, the multi-head self-attention mechanism, applied to the encoder, is also effective for long-range dependencies capturing and further semantic representation learning. The experimental results demonstrate that ALRL outperforms current existing methods in both speech quality and intelligibility, especially for low signal-to-noise ratio scenarios. Our experiments have shown that the latent space is effective to learn semantic representation with adversarial training and our ALRL is effective for speech enhancement performance improvement.

In future work, we will improve our encoder architecture for latent representation learning. Also, our improved speech enhancement method can also be applied to noise-robust speaker identification and speech recognition.
7. References


