Speech Communication 118 (2020) 1–9

Contents lists available at ScienceDirect

Speech Communication

journal homepage: www.elsevier.com/locate/specom

Improving generative adversarial networks for speech enhancement through regularization of latent representations

Fan Yang[a,b], Ziteng Wang[a,b], Junfeng Li[a,b,*], Risheng Xia[a], Yonghong Yan[a,b,c]

[a] Key Laboratory of Speech Acoustics and Content Understanding, Institute of Acoustics, Chinese Academy of Sciences, Beijing 100190, China
[b] University of Chinese Academy of Sciences, Beijing 100049, China
[c] Xinjiang Laboratory of Minority Speech and Language Information Processing, Xinjiang Technical Institute of Physics and Chemistry, Chinese Academy of Sciences, China

ARTICLE INFO

Keywords:
Generative adversarial networks
End-to-end speech enhancement
Speech enhancement under low resources

ABSTRACT

Speech enhancement aims to improve the quality and intelligibility of speech signals, which is a challenging task in adverse environments. Speech enhancement generative adversarial network (SEGAN) that adopted a generative adversarial network (GAN) for speech enhancement achieved promising results. In this paper, a new network architecture and loss function based on SEGAN are proposed for speech enhancement. Different from most network structures applied in this field, the new network, called high-level GAN (HLGAN), uses parallel noisy and clean speech signals as input in the training phase instead of only noisy speech signals, which enables us to make full use of the information carried by the clean speech signals. Additionally, we introduce a new supervised speech representation loss, also known as high-level loss, in the middle hidden layer of the generative network. The high-level loss function is advantageous to HLGAN in speech enhancement under low signal-to-noise (SNR) environments and low-resource environments. We evaluate the performance of HLGAN over a wide range of experiments, in which our model produces significant improvements. Extensive experiments further demonstrate the generality of our model in a variety of speech enhancement cases. The issue of SEGAN losing speech components while removing noise in low SNR environments is improved. In addition, HLGAN can effectively enhance the speech signals of two low-resource languages simultaneously. The reasons for the superior performance of HLGAN are discussed.

1. Introduction

Speech enhancement aims to improve the quality and intelligibility of speech signals. Whether applied to human or machine speech, it is an effective technology for noise suppression. As the front end of a speech interaction system, it is widely used and is a crucial component in many modern devices. Traditional speech enhancement methods include spectral subtraction (Boll, 1979), Wiener filtering (Lim and Oppenheim, 1978), and subspace algorithms (Dendrinos et al., 1991; Ephraim and Van Trees, 1993). Recently, deep learning-based speech enhancement methods have been widely adopted and investigated. Xu et al. trained a deep neural network (DNN) using log-power spectral features to remove noise (Xu et al., 2015; 2013). Wang et al. used a DNN to estimate masks from noisy speech signals (Wang and Chen, 2018; Li and Sim, 2014). In addition, some more complex network structures have been explored in the literature. Park and Lee (2016) employed convolutional neural networks (CNNs) for speech enhancement because CNNs have the advantage of efficiently modelling local information. Most deep learning-based methods use magnitude spectrograms as neural network inputs, which discards the phase in the speech signal. However, some recent studies have demonstrated the importance of signal phase to speech quality (Paliwal et al., 2011). Williamson et al. considered phase information using complex spectrograms (Williamson et al., 2016; Fu et al., 2017a), and this method provided a satisfactory result.

Using raw waveforms as input is another method for considering signal phase, which has the following advantages. First, it may be better at retaining speech components because there is no information loss caused by feature engineering. Second, it allows the phase embedded in the temporal domain to participate in neural network training so that the phase can be fully utilized to further improve performance. Additionally, in speech recognition, operating on the waveforms instead of MFCC features has achieved better results (Fu et al., 2017b; Palaz et al., 2013; 2015; Golik et al., 2015). However, in speech enhancement, this input approach has not been well investigated, the reason for which may be that it is more difficult to learn speech patterns from the raw waveforms than from the spectra.

* Corresponding author at: Key Laboratory of Speech Acoustics and Content Understanding, Institute of Acoustics, Chinese Academy of Sciences, Beijing 100190, China.
E-mail address: lijunfeng@hccl.ioa.ac.cn (J. Li).

https://doi.org/10.1016/j.specom.2020.02.001
Received 17 December 2018; Received in revised form 2 November 2019; Accepted 5 February 2020
Available online 6 February 2020

More recently, SEGAN has been used to perform end-to-end speech enhancement on the raw waveform and achieved promising results, which were realized by adversarial training. As with a common GAN, it consists of two components: a generator and a discriminator. The generator within SEGAN maps noisy speech signals from one distribution to enhanced speech signals from another distribution. The discriminator is in essence a classifier that labels the enhanced speech signals as false and the clean signals as true, which provides information to the generator on what to correct to produce more realistic results. In addition, the performance of SEGAN in speech enhancement under low-resource environments by exploiting transfer learning has been reported in Pascual et al. (2018).

In this study, we propose a new network architecture and a new loss function based on SEGAN and build an end-to-end speech enhancement system, which we call high-level loss generative adversarial network (HLGAN). Specifically, most neural networks applied for speech enhancement take only the spectra or masks of noisy speech signals as input without considering clean speech signals, which may lead to insufficient use of the information carried by the clean speech signals. For this, we feed a noisy speech signal and its clean equivalent into HLGAN in a parallel manner. By taking advantage of the structural characteristics of the CNN, we realize the sharing of network parameters between the noisy speech signal and the clean speech signal, which also ensures that the convolution operations performed on them are independent and identical and is different from the structures of common neural networks. In the middle hidden layer of the generative network, we obtain the corresponding outputs of the clean speech signals and noisy speech signals and take the distance between them as an optimization target, which we term high-level loss. The new network structure enables us to make full use of the information carried by clean speech signals. The new loss allows us to obtain a more accurate speech feature representation from a noisy speech signal and improves the optimization direction of the network, which boosts the speech quality and better preserves speech components while removing noise in a low SNR environment. Additionally, it benefits speech enhancement under low resources. We consider speech enhancement in different situations and implement extensive experiments to verify the validity of our model. In a wide range of experiments, our model produces significant improvements, which further illustrates that our model has generality in speech enhancement in different situations.

We organize the paper as follows: Section 2 introduces GANs and SEGAN. Section 3 details our proposed HLGAN. In Section 4, we report and analyze the experimental setups and results. Section 5 concludes our findings.

2. Speech enhancement generative adversarial network

GANs belong to generative models that attempt to map one prior distribution to another and were first proposed by Goodfellow et al. (2014). Achieving striking successes in image processing, GANs have attracted increasing attention and, hence, have various versions (e.g., deep convolutional GAN (DCGAN) (Radford et al., 2015) and Wasserstein GAN (WGAN) (Arjovsky et al., 2017)). Although GANs have been successfully applied to many problems (e.g., image generation and image classification), studies on GANs for speech-related tasks were not proposed until recently. Donahue et al. explored the effectiveness of speech enhancement with GAN for speech recognition (Donahue et al., 2018). Kim et al. proposed an acoustic and adversarial supervision approach to overcome the limitations of using clean speech as the target (Kim et al., 2019). In addition, conditional GANs and SEGAN have been investigated for speech enhancement in Michelsanti and Tan (2017) and Pascual et al. (2017).

GANs are composed of two components, a generator or generative network (G network), and a discriminator or discriminative network (D network), that compete with each other. The two components can be viewed as players in a two-player game. The discriminator attempts to distinguish between samples drawn from the training data and samples drawn from the generator, while the generator attempts to fool the discriminator into believing its samples are real (LeCun et al., 2015). The competitive relationship continuously optimizes their respective network parameters and ultimately minimizes the statistical divergence between the model distribution and the real data distribution.

In the GAN framework, we define x as a real data sample drawn from a distribution, $p_{data}(x)$. z is a random vector from distribution p_z. The G network takes z as input and outputs a generated sample, $G(z)$. The D network outputs a scalar, which represents the probability that an input sample is drawn from the distribution $p_{data}(x)$. The training process can be described as the following equation:

$$\min_G \max_D V(D,G) = E_{x\sim p_{data}(x)}[log D(x)] + E_{z\sim p_z(z)}[log(1 - D(G(z)))] \quad (1)$$

SEGAN utilizes the characteristic that GANs continuously reduce the gap between the model distribution and the real data distribution during training to enhance distorted speech signals on the raw waveform. The generative network is depicted in Fig. 1(a). In this case, the G network within SEGAN performs the speech enhancement task and adopts the encoder-decoder framework, which is divided into two parts: an encoder and a decoder. From this point, speech enhancement can be regarded as a sequence-to-sequence task, which converts one input sequence to another. The encoder is composed of strided convolutional layers, and every layer is followed by parametric rectified linear units (PReLUs). The encoder compresses its input, a noisy speech signal $\tilde{x}$, into a specific code h. The specific code h and the latent random vector z, which obeys the distribution $p_z(z)$, are concatenated along the channel dimension as the decoder input. The decoder and the encoder have a symmetrical structure. With blocks of deconvolutions, the decoder recovers a clean version of the speech from the specific code h. In addition, the G network has skip connections (He et al., 2016) that connect each encoding layer to its matching decoding layer. The skip connection attempts to solve the vanishing gradients problem, which is important for training deep neural networks because it makes the gradients flow deeper through the whole neural network (Pascual et al., 2017). The D network determines which samples are clean speech and which samples are enhanced speech. The G network uses this feedback information to correct its output.

In SEGAN, Pascual et al. utilize a conditional version of GANs to perform mapping and classification (Pascual et al., 2017; Isola et al., 2017). Meanwhile, conventional GANs often use the sigmoid cross entropy loss function in the D network. However, in Mao et al. (2017), Mao et al. think that it may result in the vanishing gradients problem. Therefore, SEGAN adopts the least-squares GAN (LSGAN) approach (Mao et al., 2017). Furthermore, to generate a more realistic sample, the L1 norm is used to measure the distance between an enhanced speech signal $G(z,\tilde{x})$ and a clean signal x. $p_{data}(x)$ and $p_{data}(\tilde{x})$ represent the distribution of clean speech signals and noisy speech signals, respectively. The loss function of SEGAN is:

$$\min_G V_{SEGAN}(G) = \frac{1}{2}E_{z\sim p_z(z),\tilde{x}\sim p_{data}(\tilde{x})}[(D(G(z,\tilde{x}),\tilde{x}) - 1)^2] + \lambda\|G(z,\tilde{x}) - x\|_1 \quad (2)$$

$$\min_D V_{SEGAN}(D) = \frac{1}{2}E_{x,\tilde{x}\sim p_{data}(x,\tilde{x})}[(D(x,\tilde{x}) - 1)^2] + \frac{1}{2}E_{z\sim p_z(z),\tilde{x}\sim p_{data}(\tilde{x})}[D(G(z,\tilde{x}),\tilde{x})^2] \quad (3)$$

Finally, SEGAN successfully applies GANs to speech enhancement.

3. High-level generative adversarial network

In SEGAN, the G network maps the noisy speech signals to the enhanced version and adopts the encoder-decoder framework. The encoder compresses its input through multiple strided convolutional layers, obtaining an output result from each layer. This compression process is performed until we obtain the vector h from the output of the

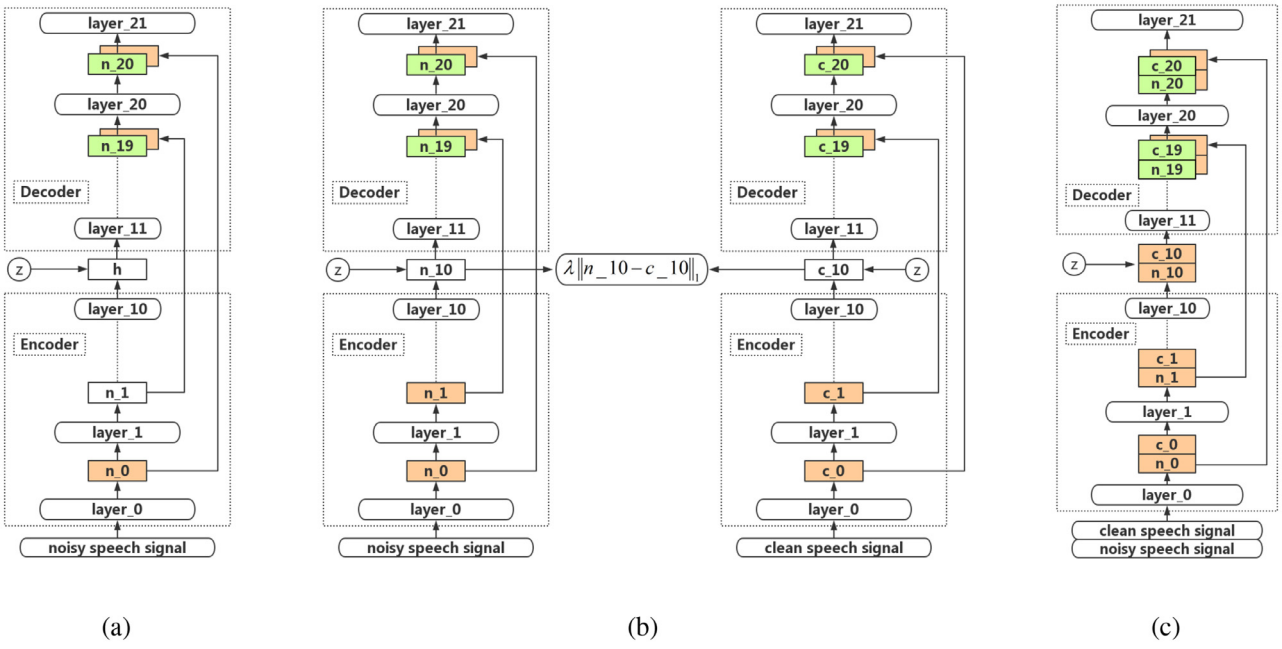

Fig. 1. (a) The G network structure of SEGAN (b) The unrolled G network structure of HLGAN (c) The actual G network structure of HLGAN.

encoder. The vector h concatenated with a random vector z is input to the decoder, which recovers the clean speech version by a series of deconvolutions. Thus, we consider that vector h is the feature representation of the speech components in speech signals and should contain all the speech information. Therefore, vector h is important for the decoder to recover speech signals. If we can obtain a better h, then we should be able to boost the speech quality of the enhanced speech signals.

Motivated by this idea, we consider obtaining a better h using supervised learning. Therefore, we feed a clean speech signal into the encoder and obtain its feature representation as a correct label. We take the distance between the label and the feature representation of the noisy speech signal as part of the optimization objective, forcing the encoder to learn an h closer to the label. This is the intended goal. This process can be described as follows. We use two identical models to separately process the noisy speech signal and the clean counterpart, obtain the corresponding feature representation and integrate the distance between the two feature representations into the loss function to train the neural network. This process is shown in Fig. 1(b), which depicts an unrolled G network structure of HLGAN.

However, implementing this idea requires meeting a set of conditions. First, this implementation requires the same parameters to process the noisy and clean speech signals. However, during the training process, the model parameters are updated in each back-propagation; thus, we must ensure that the two speech signals are processed simultaneously and that the convolution operations on both signals are the same. Second, we must ensure that these two signals are compressed independently. More importantly, in the training stage, we use clean speech signals as a reference, which, are hidden in the tests, so we must ensure that the model produced by the training can be used for testing. We propose HLGAN to solve the above problems. The G network structure of HLGAN is shown in Fig. 1(c).

Actually, we feed a noisy speech signal and the clean counterpart into the G network in parallel. Specifically, we splice noisy speech signals and clean speech signals in the width dimension and then obtain a two-dimensional vector as the input of HLGAN network, as shown in Fig. 1(c). By using convolution kernels, which are one-dimensional and have a stride of 1 along the width dimension, we guarantee meeting the above conditions. Through multi-layer convolution operations, we can obtain the feature representation n_10 of the noisy speech signal and the feature representation c_10 of the clean speech signal from the G network encoder output, as shown in Fig. 1(c). Then, we use the $L1$ norm to measure the distance between n_10 and c_10, which can be represented by the following equation:

$$Loss_{feature} = \lambda\|n\_10 - c\_10\|_1 \tag{4}$$

λ is a weight factor. Given the loss function $Loss_{feature}$, the final loss function of the G networks can be represented as the following equation:

$$\min_G V_{HLGAN}(G) = \frac{1}{2}E_{z\sim p_z(z),\tilde{x}\sim p_{data}(\tilde{x})}[(D(G(z,\tilde{x}),\tilde{x})-1)^2] + Loss_{feature} + \lambda\|G(z,\tilde{x})-x\|_1 \tag{5}$$

In this way, we realize our idea and introduce supervised speech feature representation loss, also known as high-level loss, to HLGAN. The high-level $Loss_{feature}$ creates the following benefits:

In low SNR environments, noise has much more energy than speech signals and, hence, completely covers the speech pattern in the time domain signal, and thus, the neural network cannot learn the speech feature representation; therefore, this leads to an incomplete h. When the decoder tries to reconstruct the clean speech signal by h, some speech frequency components may be lost. However, in HLGAN, we use clean speech signal waveforms as a reference. The neural network should be able to learn the speech feature representation corresponding to the low SNR regime from the reference signal, which is exactly what vector h is missing in SEGAN. This forces the encoder to learn speech feature representations from low SNR regime. Additionally, each layer of the encoder extracts and combines features on the output of the previous layer, which makes its output feature representation more abstract. With layer-by-layer convolution operations, we obtain high-level features, vector h, from low-level features. Vector h may be the fundamental properties of the input speech signal and contains abstract features of each phoneme. In SEGAN, since there is no constraint on h, it is not guaranteed that h will be accurate after multi-layer convolution operations. In other words, h may contain noise information, which may distort phonemes abstract features. However, in HLGAN, we take the high-level $Loss_{feature}$ as part of the optimization objective, which makes the n_10 as similar to

Table 1
Parameters of the G network of HLGAN.

	layer	kernel	output size ([height, width, feature maps])		layer	kernel	output size ([height, width, feature maps])
Encoder	layer_0	16	[8192, 2, 16]	Decoder	layer_11	512	[16, 2, 512]
	layer_1	32	[4096, 2, 32]		layer_12	256	[32, 2, 256]
	layer_2	32	[2048, 2, 32]		layer_13	256	[64, 2, 256]
	layer_3	64	[1024, 2, 64]		layer_14	128	[128, 2, 128]
	layer_4	64	[512, 2, 64]		layer_15	128	[256, 2, 128]
	layer_5	128	[256, 2, 128]		layer_16	64	[512, 2, 64]
	layer_6	128	[128, 2, 128]		layer_17	64	[1024, 2, 64]
	layer_7	256	[64, 2, 256]		layer_18	32	[2048, 2, 32]
	layer_8	256	[32, 2, 256]		layer_19	32	[4096, 2, 32]
	layer_9	512	[16, 2, 512]		layer_20	16	[8192, 2, 16]
	layer_10	1024	[8, 2, 1024]		layer_21	1	[16384, 2, 1]

the c_10 as possible. When n_10 is equal to c_10, the encoder can extract the high-level feature representation that contains only speech information from the noisy speech signal. This process actually removes noise in the high-level feature representation. From the above two aspects, our model has advantages in a low SNR environment.

4. Experimental setup

4.1. Dataset

To compare the performance of SEGAN with that of HLGAN, we prepare the same data for training and testing SEGAN and HLGAN in each experiment. We use the dataset by Valentini-Botinhao et al. (2016) in Section 4.4, which is the same as that in Pascual et al. (2017). In Section 4.5, the test set in Section 4.4 and four types of noise from the 100 Nonspeech Sounds corpus (Hu, 2004; Hu and Wang, 2010) are used to prepare a new test set, which is employed to evaluate the performance of the models in speech enhancement under unseen noises and unseen SNR conditions. In Section 4.6, the Korean Broadcast News Speech (*LDC*2006*S*42), 863 Chinese dataset and *Noisex*92 dataset are used to build the training set and test set to evaluate the performance of SEGAN and HLGAN in speech enhancement under low-resource environments.

4.2. Architecture and setup

In all experiments, we use the same frame-wise processing method. The original speech is sampled to 16 kHz. During the training, we use a sliding window of 16,384 points across a noisy utterance to obtain a training sample every 8192 points (50% overlap). During the test, we use the same window length, but there is no overlap between the test samples. We split one utterance to obtain multiple test samples, input every test sample to the G network, and concatenate these test sample results to obtain the final enhanced speech signal.

Regarding the structure of HLGAN, the G network has an encoder and a decoder, both of which are composed of eleven strided convolutional layers. Each layer has a different number of convolutional kernels, all of which have a height of 31, a width of 1, a stride of 2 along the height dimension, and a stride of 1 along the width dimension. In the decoder, the output of each layer and the matrix from the skip connection are combined to form the input of the next layer. Table 1 shows the number of convolutional kernels and output dimensions of each G network layer.

The structure of the D network is similar to that of the encoder, which also contains eleven strided convolutional layers. The difference between them is that the D network has two more layers than the encoder. One layer is a 1*1 convolutional layer, and the other layer is a fully connected layer. SEGAN and HLGAN have the same numbers of layers, and the numbers of convolutional kernels per layer are also consistent. However, the input of the network, the output of each convolutional layer and the step size of the convolutional kernels in the width dimension are different.

4.3. Evaluation metrics

In each experiment, we use the same objective evaluation metrics. The perceptual evaluation of speech quality (PESQ) is used to evaluate speech quality, and we adopt the wide-band version recommended in ITU-T P.862.2 (Rec, 2005). CSIG is a mean opinion score (MOS) prediction of speech signal distortion (Hu and Loizou, 2008). CBAK is an MOS prediction of the intrusiveness of background noise (Hu and Loizou, 2008). COVL is an MOS prediction of the overall effect (Hu and Loizou, 2008). SSNR is the segmental SNR (Quackenbush et al., 1988), which is fixed to the range between 35 dB and -10 dB (Papamichalis, 1987).

4.4. Performance comparison of SEGAN and HLGAN

In this experiment, the parameters of HLGAN are the same as those of SEGAN, except for epoch and effective batch size. We set 150 epochs and an effective batch size of 200. The λ of Eq. (5) is set to 100. The training set and the test set have 11,572 and 824 utterances, respectively. In the training set, there are 28 speakers, including 14 men and 14 women, ten types of noise and four SNR conditions (0 dB, 5 dB, 10 dB and 15 dB). In the test set, there are 2 speakers (a male and a female), five types of noise and four SNR conditions (2.5 dB, 7.5 dB, 12.5 dB and 17.5 dB). The five types of noise are different from those in the training set.

For comparison, we also show the results of these objective metrics applied to the noisy speech signals, the speech signals enhanced by the Wiener method based on a priori SNR estimation Loizou (2007), and the speech signals enhanced by a bidirectional long short-term memory (LSTM) model. The bidirectional LSTM has three layers, each of which has 1024 hidden units. We calculate 513-dimensional short-time Fourier transform (STFT) magnitude spectra with a 64 ms Hamming window and 32 ms overlap between frames and take it as the input feature. This model is trained for 20 epochs to predict the phase-sensitive mask (PSM) with the Adam optimizer and a learning rate of 0.0001.

From the results,[1] HLGAN achieves the best performance among all competitors on all five metrics. The PESQ score indicates that the speech quality of HLGAN is greatly improved. HLGAN produces less speech distortion in terms of CSIG and removes noise more effectively in terms of CBAK and SSNR. It also achieves a better overall effect in terms of COVL.

In addition, we compare the spectra of speech signals enhanced by SEGAN and HLGAN in low SNR environments and observe that HLGAN better preserves speech components while removing noise. We select one speech signal ($p257\_090.wav$) as an example in the test dataset. Fig. 2(a)–(d) show these spectra of the clean speech signal, the en-

[1] Some enhanced speech audio samples are provided at https://blue3sky7dream9.github.io./demo.html.

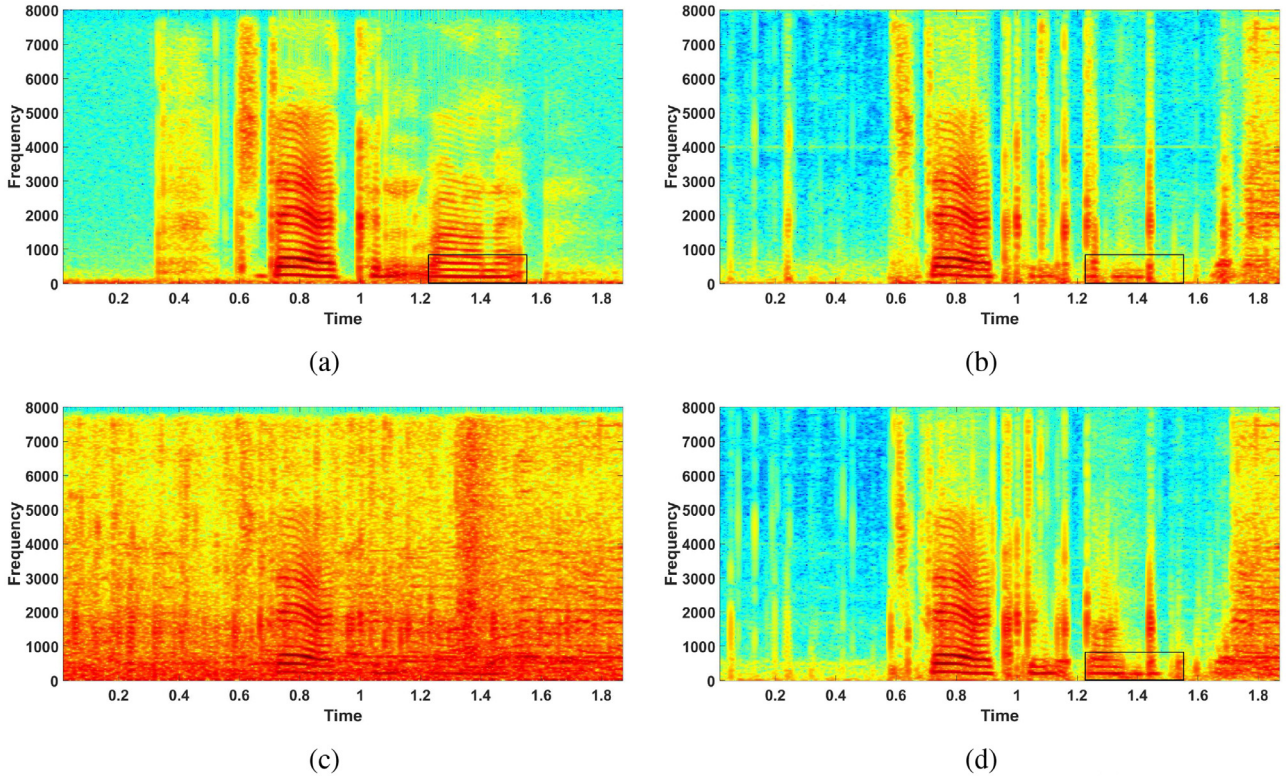

Fig. 2. Spectrograms of speech p257_090.wav in the test: (a) clean speech, (b) enhanced speech by SEGAN, (c) noisy speech, (d) enhanced speech by HLGAN .

hanced speech signal by SEGAN, the noisy speech signal and the enhanced speech signal by HLGAN. From Fig. 2(a) and b), we can observe that the speech signal enhanced by SEGAN obviously loses some speech frequency components and uses the black box to mark the corresponding spectrum, which may be because in the area where the speech component is lost, noise completely covers the speech pattern in the time domain signal. However, compared with SEGAN, HLGAN retains speech components better, which indicates that it has advantages in low SNR environments.

Then, in order to show listeners' preferences, we perform the subjective evaluation. For that, a total of 21 sentences are selected from the test set and presented to 11 listeners in a random order. All listeners are native English speakers. For each sentence, the noisy speech signal, Wiener-enhanced speech signal, BLSTM-enhanced speech signal, SEGAN-enhanced speech signal, and HLGAN-enhanced speech signal are presented to the listeners in random order. The listeners rate the overall speech quality of each speech signal on a scale of 1 to 5. According to the scoring description, they are asked to pay attention to both the signal distortion and the noise intrusiveness (e.g., 5=excellent: very natural speech with no degradation and not noticeable noise).

Table 3 reports the subjective score of each model, which shows how HLGAN is preferred over other models. Further, by subtracting the MOS of the two comparison models, we compute the comparative MOS (CMOS). Fig. 3 shows a box plot of the comparative MOS. From the figure, we can see how HLGAN is preferred. Compared with noisy speech signals, HLGAN's speech signals are preferred in 80% of cases, noisy speech signals are preferred in 9% of cases, and neither is preferred in 11% of cases. Compared with the Wiener, HLGAN is preferred in 81% of cases, the Wiener is preferred in 7% of cases, and neither is preferred in 12% of cases. Compared with the BLSTM system, HLGAN is preferred in 69% of cases, BLSTM is preferred in 15% of cases, and neither is preferred in 16% of cases. Compared with SEGAN, HLGAN is preferred in 69% of cases, SEGAN is preferred in 13% of cases, and neither is preferred in 18% of cases.

To further explain the reasons for the excellent performance of HLGAN, we compare the G network loss functions of SEGAN and HLGAN. The difference between SEGAN and HLGAN is the additive $Loss_{feature}$. The loss function of the G network within SEGAN is shown in Eq. (2). In this equation, $\lambda||G(z,\tilde{x})-x||_1$ is the distance between the enhanced speech signal and the clean signal, which is represented by $Loss_{wave}$.

$$Loss_{wave} = \lambda||G(z,\tilde{x})-x||_1 \tag{6}$$

During training, we take Eq. (2) as the optimization objective, which has two elements. One element comes from the D network, and its value is not directly related to the $Loss_{feature}$. The other element is the $Loss_{wave}$, and the $Loss_{feature}$ has a direct impact on it. However, minimizing the $Loss_{wave}$ does not guarantee an improvement in speech quality (Fu et al., 2018). For example, removing the noise in silent regions can reduce the $Loss_{wave}$, but the improvement in speech quality is very limited. The decrease in the $Loss_{wave}$ may be due to an improvement in speech quality or to the removal of noise from the silent regions. We believe that the additive $Loss_{feature}$ has a constraint on the optimization direction of the $Loss_{wave}$. Specifically, the $Loss_{feature}$ makes the $Loss_{wave}$ more likely to be optimized in the direction of improving speech quality. Thus, the correlation between the $Loss_{wave}$ and speech quality in HLGAN should be greater than that in SEGAN. To prove this idea, we perform an analysis of the test set. We process the speech signals according to the frame-wise processing method. We define that $\widetilde{c}_j$ represents a clean utterance in the test set, and $\widetilde{e}_j$ represents the enhanced utterance corresponding to $\widetilde{c}_j$. We slide the 16,384 point window across $\widetilde{c}_j$ to obtain m examples $(c\_1, c\_2, \ldots\ldots, c\_m)$, each example is 16,384 points in length. We perform the same operations on $\widetilde{e}_j$ and obtain m examples $(e\_1, e\_2, \ldots\ldots, e\_m)$. Then, we calculate the following expression on each utterance in the test set:

$$d_j = \frac{1}{m}\Sigma_{i=1}^{m}\lambda||e\_i - c\_i||_1 \tag{7}$$

$$p_j = PESQ(\widetilde{c}_j, \widetilde{e}_j) \tag{8}$$

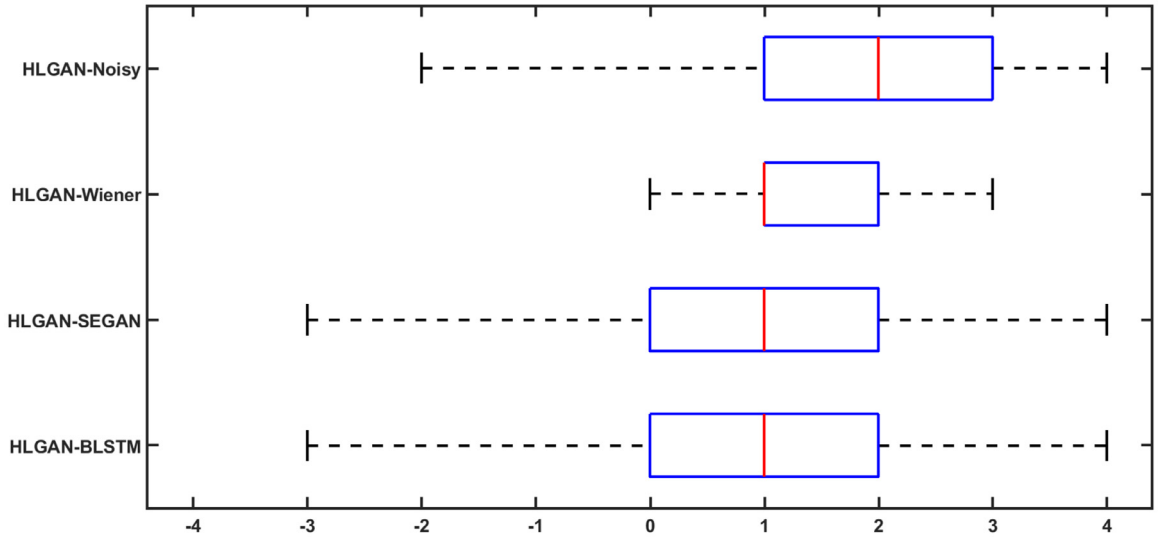

Fig. 3. CMOS box plot. Positive values mean that HLGAN is preferred.

The value of λ is 100. We obtain the two vectors *delta* and *pesq* ($delta = \{d_1, d_2 \ldots., d_{824}\}$, $pesq = \{p_1, p_2 \ldots., p_{824}\}$[2]) and then compute the correlation coefficient between them.

$$\rho = \frac{cov(delta, pesq)}{\sigma_{delta}\sigma_{pesq}} \tag{9}$$

In the above formula, $cov(delta, pesq)$ represents the covariance between the vector *delta* and the vector *pesq*, σ_{delta} and σ_{pesq} represent the standard deviation of the *delta* and *pesq*, respectively. The ρ_{SEGAN} is -0.52, and the ρ_{HLGAN} is −0.62. The greater the absolute value of the ρ, the greater the correlation between the two vectors is.

The d_j of Eq. (7) is the average of $Loss_{wave}$ in each utterance. The correlation coefficient between the $Loss_{wave}$ and PESQ is consistent with that between *delta* and *pesq*. The absolute value of the ρ of HLGAN is 0.1 higher than that of SEGAN. Therefore, we propose that the correlation between the $Loss_{wave}$ and PESQ in HLGAN is higher than that in SEGAN, which is caused by the high-level $Loss_{feature}$ in Eq. (5). The high-level $Loss_{feature}$ in HLGAN increases the correlation between the $Loss_{wave}$ and PESQ, which indicates that the decline in the $Loss_{wave}$ loss function in HLGAN is more likely to lead to an increase in the PESQ score compared to that in SEGAN, which proves our idea and explains why adding the $Loss_{feature}$ to the loss function of the G network can improve speech quality.

We also calculate the average of the $Loss_{wave}$ and $Loss_{feature}$ on all utterances in the test set, which are 645.26 and 4.73e-03 in SEGAN and 528.75 and 1.29e-08 in HLGAN. We can conclude that optimizing only the $Loss_{wave}$ in SEGAN does not guarantee a smaller $Loss_{feature}$. Simultaneously optimizing both the $Loss_{feature}$ and $Loss_{wave}$ allows us to obtain a smaller $Loss_{feature}$, which means a more accurate feature representation, while further reducing the $Loss_{wave}$. This may mean that the high-level $Loss_{feature}$ can make HLGAN converge to a better local minimum.

4.5. Testing with unseen noises and unseen SNR conditions

In Section 4.4, all the SNR conditions are greater than 0 dB. The results reported in Table 2 are the average scores of all utterances in the test set. The second experiment is designed to evaluate HLGAN and SEGAN under unseen SNR and unseen noise conditions to show their robustness. It is worth noting that in this experiment, we directly apply the models trained in Section 4.4 to process the speech signals of the test set without any training process. Fifteen clean utterances for each speaker in the test set of Section 4.4 are randomly selected and mixed with four unseen noises (N20, N27, N46, and N73) from the Nonspeech 100 Sounds corpus (Hu, 2004). N20 and N27 belong to machine noise. N46 belong to traffic and car noise, and N73 is wind noise. We consider four SNR levels (-10 dB, -7 dB, -5 dB, -3 dB) and report the results in Table 4. We use bold fonts to represent the best result of the five, and bold italics to represent the case in which the result is worse than that of the noisy speech signals.

Table 2
Objective evaluation results of different models.

Metric	Noisy	Wiener	BLSTM	SEGAN	HLGAN
PESQ	1.97	2.22	2.33	2.16	**2.48**
CSIG	3.35	3.23	3.60	3.48	**3.65**
CBAK	2.44	2.68	3.12	2.94	**3.19**
COVL	2.63	2.67	2.95	2.80	**3.05**
SSNR	1.68	5.07	9.16	7.73	**9.21**

Table 3
Subjective evaluation results of different models.

Metric	Noisy	Wiener	BLSTM	SEGAN	HLGAN
MOS	2.42	2.64	3.34	3.23	**4.09**

The table provides the following observations. 1) Compared with the results of the noisy speech signals, the PESQ scores of Wiener, BLSTM, SEGAN and HLGAN are all improved. However, the improvements of HLGAN are the most obvious. 2) At −10 dB, −7 dB and −5 dB, the CSIG scores of Wiener are slightly lower than those of the noisy speech signals. At −10 dB, SEGAN has similar performance with the noisy speech signals in terms of CSIG. 3) Under unseen noise types and worse SNR conditions than those of the training set, even at −10 dB, the PESQ score of HLGAN is improved. Meanwhile, a maximum gain of 22.73% relative to the noisy speech signals is obtained at −3 dB. 4) HLGAN achieves the best performance on all five metrics under all SNR conditions.

Additionally, we reported the results of the objective metrics on the four types of noise in Table 5.

As shown in Table 5, we observe the following. 1) The PESQ scores of HLGAN have different gains on different noise types, and the maximum

[2] We have 824 delta and pesq values since there are 824 utterances in the test set.

Table 4
Objective results comparing the noisy signals and Wiener- and BLSTM- and SEGAN- and HLGAN-enhanced signals under all SNR conditions.

	PESQ					CSIG					CBAK					COVL					SSNR				
SNR(dB)	Noisy	Wiener	BLSTM	SEGAN	HLGAN	Noisy	Wiener	BLSTM	SEGAN	HLGAN	Noisy	Wiener	BLSTM	SEGAN	HLGAN	Noisy	Wiener	BLSTM	SEGAN	HLGAN	Noisy	Wiener	BLSTM	SEGAN	HLGAN
-10	1.05	1.07	1.11	1.09	**1.12**	1.68	***1.62***	***1.58***	***1.67***	**1.79**	1.09	1.14	1.41	1.48	**1.53**	1.24	***1.21***	***1.23***	1.26	**1.34**	-8.61	-6.90	-3.49	-2.71	**-2.30**
-7	1.06	1.10	1.17	1.14	**1.19**	1.83	***1.80***	1.85	1.93	**2.09**	1.18	1.24	1.62	1.66	**1.76**	1.33	***1.32***	1.40	1.43	**1.55**	-7.66	-5.87	-1.80	-1.08	**-0.32**
-5	1.07	1.12	1.22	1.18	**1.27**	1.95	***1.94***	2.04	2.12	**2.29**	1.26	1.34	1.76	1.79	**1.91**	1.40	1.40	1.54	1.56	**1.70**	-6.90	-5.07	-0.65	-0.03	**0.90**
-3	1.10	1.16	1.29	1.26	**1.35**	2.07	2.08	2.26	2.32	**2.47**	1.35	1.44	1.90	1.93	**2.06**	1.48	1.51	1.70	1.71	**1.84**	-6.05	-4.19	0.47	1.00	**2.04**

Table 5
Objective results comparing the noisy signals and Wiener- and BLSTM- and SEGAN- and HLGAN-enhanced signals on four types of noise.

	PESQ					CSIG					CBAK					COVL					SSNR				
	Noisy	Wiener	BLSTM	SEGAN	HLGAN	Noisy	Wiener	BLSTM	SEGAN	HLGAN	Noisy	Wiener	BLSTM	SEGAN	HLGAN	Noisy	Wiener	BLSTM	SEGAN	HLGAN	Noisy	Wiener	BLSTM	SEGAN	HLGAN
N20	1.05	1.05	**1.15**	1.11	**1.15**	1.88	***1.86***	1.97	1.91	**2.03**	1.18	***1.17***	1.54	1.58	**1.66**	1.34	***1.33***	1.44	1.40	**1.49**	-7.53	-7.46	-2.43	-1.97	**-1.19**
N27	1.05	1.05	**1.15**	1.11	**1.15**	1.85	***1.83***	1.98	1.90	**2.02**	1.17	1.17	1.54	1.58	**1.66**	1.33	***1.32***	1.44	1.40	**1.48**	-7.56	-7.50	-2.42	-1.96	**-1.20**
N46	1.11	1.28	1.36	1.34	**1.49**	2.18	2.28	2.27	2.49	**2.70**	1.35	1.64	2.07	2.12	**2.27**	1.56	1.67	1.75	1.85	**2.04**	-6.80	-2.29	2.02	2.65	**3.58**
N73	1.07	1.07	1.14	1.11	**1.15**	1.63	***1.47***	***1.51***	1.74	**1.90**	1.17	1.18	1.54	1.59	**1.66**	1.22	***1.12***	1.23	1.31	**1.41**	-7.34	-4.79	-2.62	-1.54	**-0.87**

Table 6
Performance comparison of different models before and after transfer learning for Korean.

Korean	PESQ					CSIG					CBAK					COVL					SSNR				
	Noisy	Wiener	BLSTM	SEGAN	HLGAN	Noisy	Wiener	BLSTM	SEGAN	HLGAN	Noisy	Wiener	BLSTM	SEGAN	HLGAN	Noisy	Wiener	BLSTM	SEGAN	HLGAN	Noisy	Wiener	BLSTM	SEGAN	HLGAN
Before	1.21	**1.47**	***1.18***	***1.11***	***1.12***	1.59	**1.82**	***1.36***	***1.56***	***1.57***	2.27	**2.43**	***2.26***	***2.01***	***2.03***	1.38	**1.60**	***1.24***	***1.28***	***1.28***	5.05	**7.23**	5.47	***3.18***	***3.62***
After	1.21	1.47	1.40	1.54	**1.65**	1.59	1.82	2.06	2.25	**2.42**	2.27	2.43	2.65	2.58	**2.66**	1.38	1.60	1.70	1.86	**2.01**	5.05	7.23	**9.25**	7.29	7.53

gain is obtained on N46. 2) On N20, N27, and N73, Wiener performs similarly and worse than with the noisy speech signals in terms of PESQ, CSIG and COVL, which indicates that Wiener cannot enhance the noisy speech signals very well. 3) On all noise types, all the objective evaluation scores of HLGAN are better than those of the noisy speech signals, Wiener, BLSTM and SEGAN, except that HLGAN and BLSTM have the same PESQ scores on N20 and N27.

In this section, we consider speech enhancement under unseen noises and unseen SNR conditions. We built a new test set to evaluate the robustness of our model using unseen noises and worse SNR environments than the training set. In all SNR conditions, HLGAN has the best performance on all metrics. On all unseen noise types, HLGAN achieves a better enhancement effect. These results indicate that HLGAN has better robustness in speech enhancement under unseen noises and unseen SNR conditions.

4.6. Speech enhancement under low-resource conditions

Deep learning-based speech enhancement methods often require large quantities of training data. However, for many new and low-resource languages, it is very difficult to obtain these large-scale data, which makes speech enhancement under low-resource conditions an important topic. Speech enhancement under low-resource conditions often adapts a model trained on a source language for a target language with transfer learning. Pascual et al. show the performance of SEGAN in this aspect (Pascual et al., 2018).

To explore whether our model can benefit speech enhancement under low-resource conditions and compare the performance with SEGAN under the same adaptation data, we conduct the following experiment.

In this experiment, we use the HLGAN and BLSTM models trained in Section 4.4 as pre-trained models of HLGAN and BLSTM and use the SEGAN model provided by the SEGAN author as a pre-trained SEGAN model. We first realize speech enhancement under low resources by transfer learning from English to Korean. The Korean Broadcast News Speech (*LDC2006S42*) and *Noisex92* datasets are used to prepare the training and test sets. The Korean Broadcast News Speech (*LDC2006S42*) dataset consists of 18 audio files recorded by the Linguistic Data Consortium (LDC) from the Voice of America (VOA) satellite radio news broadcasts in Korean. All files are split according to the transcripts provided by the LDC, resulting in a total of 4454 utterances. We randomly divide the *Noisex92* dataset into two parts, one containing 10 types of noise for building the training set and the other containing 5 types of noise for building the test set. For the training set, we randomly select 30 clean utterances in the Korean Broadcast News Speech (*LDC2006S42*) dataset and add ten types of noise to the clean speech signals at four SNR levels (0 dB, 5 dB, 10 dB, and 15 dB). For the test set, we randomly select another 20 clean utterances in the Korean Broadcast News Speech (*LDC2006S42*) dataset and add five other types of noise to the clean speech signals at four SNR levels (2.5 dB, 7.5 dB, 12.5 dB, and 17.5 dB). Considering the real environments, we make the noises and SNR conditions of the training set not match those of the test set.

We use the training set to train new models based on pre-trained models and evaluate the performances of the new models on the test set. We make the parameters of HLGAN and SEGAN consistent. To compare the results before and after the transfer learning, we use the pre-trained models to directly process the speech signals to obtain the enhanced version. The objective metric scores of the enhanced speech signals are taken as a reference. The experimental results are shown in Table 6.

From Table 6, we observe that before transfer learning, HLGAN and SEGAN perform similarly, whose performances of the two models are worse than that of the noisy speech signals. The scores of BLSTM are lower than those of the noisy speech signals in terms of PESQ, CSIG, CBAK and COVL. After transfer learning, the performance of HLGAN is greatly improved, and all the objective metric scores of HLGAN are higher than those of SEGAN, which shows that HLGAN has better per-

Table 7
The amount of data for each language.

Size(min)	English	Korean	Chinese
Train	563	139	140
Test	34	57	24

formance and proves that high-level loss benefits speech enhancement under low-resource conditions. In addition to SSNR, HLGAN has a higher score than BLSTM on all metrics.

Furthermore, we consider a more complex situation: simultaneously enhancing the speech signals of two low-resource languages.

In this experiment, we perform speech enhancement under low resources by transfer learning from one language (English) to two low-resource languages (Korean and Chinese). The 863 Chinese dataset is used to construct the Chinese set in the experiment. The training set in the 863 Chinese dataset contains 152 speakers (76 men and 76 women), and the test set contains 14 speakers (7 men and 7 women). Each speaker has approximately 30 minutes of recording. We use the Korean set constructed in the previous experiment and construct the Chinese set in the same way. The noise types and SNR levels in the Chinese training and test sets are consistent with those of the Korean set. For the training set, 64 clean utterances are selected from the 863 Chinese dataset and corrupted with ten types of noise at four SNR levels (0 dB, 5 dB, 10 dB, and 15 dB). For the test set, 20 clean utterances are excerpted from the test set of the 863 Chinese dataset and corrupted with five types of noise at four SNR levels (2.5 dB, 7.5 dB, 12.5 dB, and 17.5 dB). We report the amount of data for each language in Table 7. We mix the Chinese training set and the Korean training set into a new training set for training. All the pre-trained models in the experiment are the same as those in the previous experiment. We train SEGAN, BLSTM and HLGAN with the new training set and evaluate their performance on the Chinese test set and the Korean test set respectively. In the training stage, SEGAN and HLGAN parameters are consistent. Table 8 shows the results for Chinese and Table 9 shows the results for Korean.

From Table 8, we observe that when we directly apply the pre-training models of SEGAN, BLSTM and HLGAN to the Chinese dataset, their scores are lower than those of Wiener in terms of PESQ, CBAK, COVL and SSNR. After transfer learning, HLGAN achieves a significant improvement. Meanwhile, all scores of HLGAN are higher than those of Wiener. However, the PESQ score of SEGAN is lower than that of Wiener, which may illustrate that the enhancement effect of SEGAN is limited. In addition, HLGAN outperforms SEGAN and BLSTM on all objective evaluation metrics.

From the results of Table 9, before transfer learning, SEGAN, BLSTM and HLGAN do not perform well on the Korean set. Their performances are worse than those of Wiener. After transfer learning, the scores of SEGAN and HLGAN are higher than those of Wiener. The performances of SEGAN and HLGAN are significantly improved. However, HLGAN has a better performance.

In this section, we consider speech enhancement under low-resource conditions. To evaluate whether our model benefits speech enhancement under low resources, we compare the performance of SEGAN and HLGAN on the Korean data set. The experimental results show that our model can better enhance the speech signals of the low-resource language. Furthermore, we consider a more complex case: enhancing the speech signals of two low-resource languages at the same time. In this experiment, we observe that when SEGAN and HLGAN enhance the low-resource speech signals, HLGAN can achieve better performance than SEGAN under the same conditions, which indicates that our model can better improve the speech enhancement quality under low-resource environments. The results of this experiment also indicate that our model can effectively enhance the speech signals of Korean and Chinese.

Table 8
Performance comparison of different models before and after transfer learning for Chinese.

Chinese	PESQ					CSIG					CBAK					COVL					SSNR				
	Noisy	Wiener	BLSTM	SEGAN	HLGAN	Noisy	Wiener	BLSTM	SEGAN	HLGAN	Noisy	Wiener	BLSTM	SEGAN	HLGAN	Noisy	Wiener	BLSTM	SEGAN	HLGAN	Noisy	Wiener	BLSTM	SEGAN	HLGAN
Before	1.32	**1.69**	1.45	1.37	1.43	2.29	2.55	**2.26**	2.60	**2.66**	1.95	**2.17**	2.10	2.00	2.14	1.74	**2.02**	1.78	1.91	1.98	0.27	**2.62**	2.10	1.12	2.36
After	1.32	1.69	1.69	1.63	**1.79**	2.29	2.55	2.78	2.70	**2.83**	1.95	2.17	2.36	2.21	**2.41**	1.74	2.02	2.18	2.10	**2.26**	0.27	2.62	3.60	2.15	**3.68**

Table 9
Performance comparison of different models before and after transfer learning for Korean.

Korean	PESQ					CSIG					CBAK					COVL					SSNR				
	Noisy	Wiener	BLSTM	SEGAN	HLGAN	Noisy	Wiener	BLSTM	SEGAN	HLGAN	Noisy	Wiener	BLSTM	SEGAN	HLGAN	Noisy	Wiener	BLSTM	SEGAN	HLGAN	Noisy	Wiener	BLSTM	SEGAN	HLGAN
Before	1.21	**1.47**	***1.18***	***1.11***	***1.12***	1.59	**1.82**	***1.36***	***1.56***	***1.57***	2.27	**2.43**	***2.26***	***2.01***	***2.03***	1.38	**1.60**	***1.24***	***1.28***	***1.28***	5.05	**7.23**	5.47	***3.18***	***3.62***
After	1.21	1.47	1.43	1.56	**1.71**	1.59	1.82	2.08	2.16	**2.32**	2.27	2.43	2.66	2.57	**2.70**	1.38	1.60	1.72	1.83	**1.99**	5.05	7.23	**9.32**	6.97	7.74

5. Conclusion

In this study, we propose a new network architecture and loss function based on SEGAN, which has advantages in speech enhancement under low SNR environments and low-resource environments. HLGAN directly processes the waveforms of speech signals, which avoids the transformation between the time domain and the frequency domain. Different from the existing neural network structures, the new network allows us to feed the clean and noisy speech signals into HLGAN in parallel, which makes full use of the information carried by the clean speech signal. The new loss function enables us to obtain a more accurate speech feature representation from a noisy speech signal and limits the optimization direction of the $Loss_{wave}$ loss function. We explained the reasons for the excellent performance of HLGAN from different angles and proved them. By comparing the correlation coefficient between the $Loss_{wave}$ and PESQ in two models, we proved that the $Loss_{feature}$ can improve the optimization direction of the $Loss_{wave}$. By calculating the average distance between the feature representations of the clean speech signals and the noisy speech signals in two models, we demonstrate that HLGAN can extract a more accurate feature representation from a noisy speech signal. In extensive experiments, we considered speech enhancement in different situations. In Section 4.4, we compared the performance of SEGAN and HLGAN on the same dataset, showing that HLGAN has a significant improvement. The issue that SEGAN loses speech components while removing noise is improved. In Section 4.5, we performed speech enhancement under unseen noises and unseen SNR conditions, which showed that HLGAN has better robustness and better performance in low SNR environments. In Section 4.6, we explored the speech enhancement performance of SEGAN and HLGAN under low-resource conditions and considered a more complex case. These experimental results demonstrated that HLGAN can benefit speech enhancement under low-resource conditions. Furthermore, HLGAN achieved significant improvements over a wide range of experiments, which shows that our model has generality in speech enhancement in a variety of situations. In future work, we will verify whether our model has a broader generality; that is, whether it benefits different network structures under the encoder-decoder framework.

Declaration of Competing Interest

We declare that we have no financial and personal relationships with other people or organizations that can inappropriately influence our work, there is no professional or other personal interest of any nature or kind in any product, service and/or company that could be construed as influencing the position presented in, or the review of, the manuscript.

References

Arjovsky, M., Chintala, S., Bottou, L., 2017. Wasserstein generative adversarial networks. In: International Conference on Machine Learning, pp. 214–223.

Boll, S., 1979. Suppression of acoustic noise in speech using spectral subtraction. IEEE Trans. Acoust. 27 (2), 113–120.

Dendrinos, M., Bakamidis, S., Carayannis, G., 1991. Speech enhancement from noise: a regenerative approach.. Speech Commun. 10 (1), 45–57.

Donahue, C., Li, B., Prabhavalkar, R., 2018. Exploring speech enhancement with generative adversarial networks for robust speech recognition. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 5024–5028.

Ephraim, Y., Van Trees, H.L., 1993. A signal subspace approach for speech enhancement. In: iCASSP, pp. 355–358.

Fu, S.W., Hu, T.Y., Yu, T., Lu, X., 2017. Complex spectrogram enhancement by convolutional neural network with multi-metrics learning. In: IEEE International Workshop on Machine Learning for Signal Processing, pp. 1–6.

Fu, S.-W., Wang, T.-W., Tsao, Y., Lu, X., Kawai, H., 2018. End-to-end waveform utterance enhancement for direct evaluation metrics optimization by fully convolutional neural networks. IEEE/ACM Trans. Audio Speech Lang.Process. 26 (9), 1570–1584.

Fu, S.W., Yu, T., Lu, X., Kawai, H., 2017. Raw waveform-based speech enhancement by fully convolutional networks. In: Asia-Pacific Signal and Information Processing Association Summit and Conference, pp. 006–012.

Golik, P., Tüske, Z., Schlüter, R., Ney, H., 2015. Convolutional neural networks for acoustic modeling of raw time signal in LVCSR. Sixteenth Annual Conference of the International Speech Communication Association.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative adversarial nets. In: Advances in Neural Information Processing Systems, pp. 2672–2680.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778.

Hu, G.,. 100 nonspeech environmental sounds, 2004. http://web.cse.ohio-state.edu/~wang.77/pnl/corpus/HuNonspeech/HuCorpus.html.

Hu, G., Wang, D., 2010. A tandem algorithm for pitch estimation and voiced speech segregation. IEEE Trans. Audio Speech Lang. Process. 18 (8), 2067–2079.

Hu, Y., Loizou, P.C., 2008. Evaluation of objective quality measures for speech enhancement. IEEE Trans. Audio Speech Lang. Process. 16 (1), 229–238.

Isola, P., Zhu, J.-Y., Zhou, T., Efros, A.A., 2017. Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1125–1134.

Kim, G., Lee, H., Kim, B.-K., Oh, S.-H., Lee, S.-Y., 2019. Unpaired speech enhancement by acoustic and adversarial supervision for speech recognition. IEEE Signal Process. Lett. 26 (1), 159–163.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521 (7553), 436.

Li, B., Sim, K.C., 2014. A spectral masking approach to noise-robust speech recognition using deep neural networks. IEEE/ACM Trans. Audio Speech Lang.Process. 22 (8), 1296–1305.

Lim, J.S., Oppenheim, A.V., 1978. All-pole modeling of degraded speech. Acoust. Speech Signal Process. IEEE Trans. 26 (3), 197–210.

Loizou, P.C., 2007. Speech Enhancement: Theory and Practice. CRC press.

Mao, X., Li, Q., Xie, H., Lau, R.Y., Wang, Z., Smolley, S.P., 2017. Least squares generative adversarial networks. In: Computer Vision (ICCV), 2017 IEEE International Conference on. IEEE, pp. 2813–2821.

Michelsanti, D., Tan, Z.-H., 2017. Conditional generative adversarial networks for speech enhancement and noise-robust speaker verification. arXiv:1709.01703.

Palaz, D., Collobert, R., Doss, M.M., 2013. Estimating phoneme class conditional probabilities from raw speech signal using convolutional neural networks. Comput. Sci..

Palaz, D., Magimai.-Doss, M., Collobert, R., 2015. Convolutional neural networks-based continuous speech recognition using raw speech signal. In: IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 4295–4299.

Paliwal, K., Wjcicki, K., Shannon, B., 2011. The importance of phase in speech enhancement. Speech Commun. 53 (4), 465–494.

Papamichalis, P.E., 1987. Practical Approaches to Speech Coding. Prentice-Hall, Inc..

Park, S. R., Lee, J., 2016. A fully convolutional neural network for speech enhancement. arXiv:1609.07132.

Pascual, S., Bonafonte, A., Serrà, J., 2017. Segan: speech enhancement generative adversarial network. In: Proc. Interspeech 2017, pp. 3642–3646.

Pascual, S., Park, M., Serrà, J., Bonafonte, A., Ahn, K.-H., 2018. Language and noise transfer in speech enhancement generative adversarial network. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 5019–5023.

Quackenbush, S., Barnwell, T., Clements, M., 1988. Objective measures of speech quality. 1988. Ramrez, J., JC Segura, C. Bentez, L. Garca, and A. Rubio." Statistical voice activity.

Radford, A., Metz, L., Chintala, S., 2015. Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv:1511.06434.

Rec, I., 2005. P. 862.2: Wideband Extension to Recommendation p. 862 for the Assessment of Wideband Telephone Networks and Speech Codecs. International Telecommunication Union, CH, Geneva.

Valentini-Botinhao, C., Wang, X., Takaki, S., Yamagishi, J., 2016. Investigating RNN-based speech enhancement methods for noise-robust text-to-speech. In: ISCA Speech Synthesis Workshop, pp. 146–152.

Wang, D., Chen, J., 2018. Supervised speech separation based on deep learning: an overview. IEEE/ACM Trans. Audio Speech Lang. Process..

Williamson, D.S., Wang, Y., Wang, D.L., 2016. Complex ratio masking for joint enhancement of magnitude and phase. In: IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 5220–5224.

Xu, Y., Du, J., Dai, L.R., Lee, C.H., 2013. An experimental study on speech enhancement based on deep neural networks. IEEE Signal Process. Lett. 21 (1), 65–68.

Xu, Y., Du, J., Dai, L.-R., Lee, C.-H., 2015. A regression approach to speech enhancement based on deep neural networks. IEEE/ACM Trans. Audio Speech Lang.Process. 23 (1), 7–19.