*INTERSPEECH 2019*
September 15–19, 2019, Graz, Austria

# Speech Enhancement with Variance Constrained Autoencoders

*D. T. Braithwaite, W. B. Kleijn*

School of Engineering and Computer Science
Victoria University of Wellington, New Zealand
{daniel.braithwaite, bastiaan.kleijn}@ecs.vuw.ac.nz

## Abstract

Recent machine learning based approaches to speech enhancement operate in the time domain and have been shown to outperform the classical enhancement methods. Two such models are SE-GAN and SE-WaveNet, both of which rely on complex neural network architectures, making them expensive to train. We propose using the Variance Constrained Autoencoder (VCAE) for speech enhancement. Our model uses a more straightforward neural network structure than competing solutions and is a natural model for the task of speech enhancement. We demonstrate experimentally that the proposed enhancement model outperforms SE-GAN and SE-WaveNet in terms of perceptual quality of enhanced signals.

**Index Terms**: speech enhancement, generative modelling, time domain, deep learning, neural networks

## 1. Introduction

Speech Enhancement (SE) is the task of improving the quality of an audio signal that has been corrupted by noise [1]. In recent years, neural networks have become an increasingly popular method for enhancing speech signals [2–12]. Many of these systems operate on the magnitude spectrum [4, 7–9] and provide a spectrogram output. The disadvantage of magnitude spectrum enhancement is that the time-domain signal must be re-synthesised from the processed spectrogram. While methods for constructing a time-domain signal from a spectrogram exist [13], the quality of the speech enhancement system is then limited by the effectiveness of the synthesis algorithm used.

Recently, several machine learning based speech enhancement systems have focused on time-domain enhancement [6, 10–12], SE-GAN [6] and SE-WaveNet [10] are two such methods which take a generative modelling approach. Generative modelling approaches to speech enhancement present a new paradigm, in which the objective is to generate convincing speech that matches the content of the original, noisy speech. This is in contrast to the classical methods, i.e., Wiener filter [14], which do not know the attributes of speech and commonly lead to audible distortions. For example, in an MMSE approach, if multiple speech signals are equally plausible, then the model will compromise between them, resulting in un-natural sounding speech. A generative paradigm will instead select one of the equally likely speech signals.

SE-GAN and SE-WaveNet are both based on generative models (Generative Adversarial Networks (GANs) [15], and WaveNet [16] respectively) and have been shown to enhance time-domain speech signals more effectively than the classical Wiener filtering based approaches [14]. Both GANs and WaveNet can be trained to produce convincing generations. However, it is not possible to control the content of what these models generate. In the context of speech enhancement, it is necessary to specify the linguistic and speaker information in the enhanced speech. Both SE-GAN and SE-WaveNet modify their underlying generative architectures to allow the linguistic and speaker information of the generated speech to be controlled. A consequence of these modifications is that the speech enhancement models are no-longer fully generative. On a more practical note, another issue with SE-GAN and SE-WaveNet is that they both rely on large and complex neural network architectures, making these models expensive to train and run. In the case of SE-GAN, a more fundamental issue is the recent work showing that GANs have unstable training dynamics, which can cause the discriminator and generator to diverge [17].

The Variational Autoencoder (VAE) [18,19] is another generative model that has been applied to frequency-domain speech enhancement [20,21]. VAE is a natural model to use for enhancing speech because of its encoder/decoder structure. In such a model, noisy speech, $\tilde{X}$, is encoded to a latent representation, $Z$, which is used to generate clean speech, $X$. The latent space learned by this model can be seen as the clean speech manifold. However, recent papers have shown that VAE does not necessarily learn a meaningful latent representation [22, 23], i.e., $\tilde{X}$ is independent of $Z$. This issue would mean that the linguistic and speaker information of generated speech could not be controlled. Therefore, we propose the Variance Constrained Autoencoder (VCAE) [24] for time-domain speech enhancement.

VCAE is an appropriate model for speech enhancement for several reasons. Firstly, VCAE is not bound to a complex neural network structure (like SE-WaveNet). Secondly, VCAE has an encoder/decoder structure with a learned latent representation, making it a natural fit for the speech enhancement problem. Moreover, unlike SE-GAN, VCAE explicitly prioritises learning a meaningful latent structure; this prevents overfitting and improves generalisation. Lastly, unlike other recently proposed improvements on VAE [25, 26], VCAE does not enforce a pre-defined prior on the distribution over the latent encodings, allowing the latent feature structure to represent the data better.

The proposed speech enhancement VCAE (SE-VCAE) minimises a Wasserstein distance between its generative distribution and the clean speech distribution. Minimising the aforementioned Wasserstein distance ensures that the enhanced speech sounds realistic. However, minimising Wasserstein distance alone does not allow for control over the content of the generated speech signals. Consequently, like SE-GAN, we additionally minimise an L1 error between the enhanced and desired signals. Since SE-VCAE optimises an L1 measure, it is also not a fully generative enhancement approach.

**Key contributions:** 1) We propose SE-VCAE (Section 3), a novel machine learning speech enhancement approach based on VCAE (introduced in Section 2), 2) our proposed system outperforms SE-GAN and SE-WaveNet (Section 4), and 3) SE-VCAEs implementation is more straightforward than both competing methods, demonstrating that complex neural network structures are not necessary to enhance speech.



http://dx.doi.org/10.21437/Interspeech.2019-1809

## 2. Variance Constrained Autoencoder

In this section, we introduce the Variance Constrained Autoencoder (VCAE) [24]. Throughout the remainder of this paper, random variables are denoted by capital letters, e.g., $X$, and their realisations as lower case letters, e.g., $x$. We abbreviate probability density functions $P(X = x) = p_X(x)$ as $p(x)$.

VCAE is a generative latent feature model, i.e., VCAE assumes that there is a set of underlying features, given by the random variable $Z$, which generate the data, $X \sim p_D(x)$, where $p_D(x)$ is the distribution for $X$ defined by the data. In general, we do not have access to $Z$, and we wish to infer the latent features as well as use them for generating data. Let $z_{dim}$ and $x_{dim}$ be the dimensionalities of $Z$ and $X$ respectively, typically, $z_{dim} << x_{dim}$. VCAE consists of encoder and decoder distributions, given by $q_\phi(z|x)$ and $p_\theta(x|z)$ respectively. Both the encoder and decoder are implemented by neural networks with parameters $\phi$ and $\theta$, respectively. Other, recent generative models [25,26] constrain the shape of $q_\phi(z)$, i.e. to be a Gaussian. In the case of VCAE, the shape of $q_\phi(z)$ is not constrained, only the variance is. This change in constraints allows for a latent structure that more naturally represents the data [24].

Recent papers have advocated for the use of deterministic encoders [25], where $z \sim q_\phi(z|x)$ has the form $z = \mu_\phi(x)$, where $\mu_\phi(x)$ is implemented by a neural network and outputs the mean of $q_\phi(z|x)$. However, it has been shown that the use of stochastic encoders can prevent overfitting [26]. Allowing the variance of $q_\phi(z|x)$ to change means it can be set advantageously with respect to optimising the objective, without regard to interpolation between latent points. A fixed variance ensures that interpolation across the latent space is well defined. We define $q_\phi(z|x)$ to have a fixed variance: $z = \mu_\phi(x) + \epsilon$, where $\epsilon \sim p_\phi(\epsilon)$ is a user-defined distribution

The VCAE architecture is shown in Figure 1. The distribution $q_\phi(Z)$ is not known, which makes sampling from the generative model difficult. However, this is not a problem for the application explored in this paper as we are interested in the latent structure, not sampling from $q_\phi(z)$.

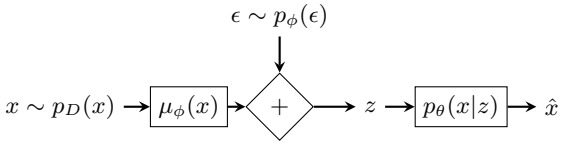



Figure 1: *VCAE Architecture.*

To train VCAE, we maximise the likelihood of the data given the corresponding latent features, subject to a constraint on the variance of the distributions $q_\phi(z)$ and $q_\phi(z|x)$. The constraint on $q_\phi(z|x)$ can be enforced directly, as $p_\phi(\epsilon)$ is fixed. However, the restriction on the variance of $q_\phi(z)$ cannot be implemented directly and requires the use of a penalty function. The objective is given by:

$$\underset{\phi,\theta}{\text{maximize}} \quad \mathbb{E}_{X\sim p_D(x)}\, \mathbb{E}_{Z\sim q_\phi(z|x;)}[\log p_\theta(X|Z)] - \lambda\, |\, \mathbb{E}_{Z\sim q_\phi(z)}[||Z - \mathbb{E}_{Z\sim q_\phi(z)}[Z]||_2^2] - v|, \quad (1)$$

where $v$ is the desired summed variance (across each dimension) of $q_\phi(z)$ and $\lambda$ is a hyperparameter. When optimising (1), $\mathbb{E}_{Z\sim q_\phi(z)}[Z]$ is not constrained. However, we do not need to constrain $\mathbb{E}_{Z\sim q_\phi(z)}[Z]$ because a shift in the mean of $q_\phi(z)$ does not affect the objective of maximising the likelihood.

## 3. VCAE Speech Enhancement Model

In this section, we first formulate the speech enhancement problem in the context of generative modelling, then we describe the Speech Enhancement Variance Constrained Autoencoder (SE-VCAE), our approach to enhancing time domain speech signals.

Let $X$ and $\tilde{X}$ be random variables representing blocks of clean and noisy speech, respectively. $X$ and $\tilde{X}$ have distribution $p_D(x)$ and $p_D(\tilde{x})$, respectively, both defined by the data. The speech enhancement problem can be formulated as learning the distribution $p(x|z)$, where $Z$ is a set of latent features that describe the clean speech being generated. We wish to learn the distribution over latent features given the noisy data, $q(z|\tilde{x})$.

This problem formulation can be implemented naturally by VCAE, with distributions $p_\theta(x|z)$ and $q_\phi(z|\tilde{x})$, we denote this model SE-VCAE. SE-VCAE differs slightly from the standard VCAE setup, where the encoder would instead be $q_\phi(z|x)$.

SE-VCAEs objective is, in part, to maximise the likelihood of clean speech given the latent vectors that were inferred from the noisy speech. In practice, computing the likelihood can be difficult. We assume that $p_\theta(x|z)$ is a factored Laplace distribution, making the log-likelihood the negative L1 error.

Ensuring that the proposed model produces convincing clean speech is an important distinction between the generative and classical enhancement paradigms. As a consequence, in addition to the standard VCAE objective, we minimise the Wasserstein distance (WD) between $p_\theta(x)$ and $p_D(x)$. It is possible, e.g., [27], to compute the WD using:

$$W_f(p_\theta(x), p_D(x)) = \sup_{||f||_L \leq 1} \mathbb{E}_{X\sim p_D(x)}[f(X)] - \mathbb{E}_{X\sim p_\theta(x)}[f(X)], \quad (2)$$

where $f$ is implemented by a neural network, of which, the Lipschitz constant ($||f||_L$) is constrained to be less than or equal to one. A method for enforcing the constraint on the Lipschitz constant of $f$ is to minimise the absolute difference between the two-norm of the gradient of $f$ and 1 [28].

It is essential that the learned model can generalise to unseen speakers and noise conditions as it is infeasible to obtain a training set with an exhaustive collection of environments. Weight regularisation is a well-established method for improving how well a model generalises [29]. Consequently, we will apply an L1 penalty to the weights of the encoder and decoder parameters. This L1 regularisation improves model generalisation to unseen noise types by reducing reliance on unnecessary features. We note that the use of stochastic encoders has a different regularising effect on the model, ensuring the latent representation is smooth. Additionally, the latent stochasticity prevents the decoder from performing poorly on latent vectors between the encoded data points.

With the addition of the Wasserstein distance and L1 regularisation, the SE-VCAEs objective function is:

$$\underset{\phi,\theta}{\text{minimise}} \quad \mathbb{E}_{X\sim p_D(x)}\, \mathbb{E}_{Z\sim q_\phi(z|x)}[||X - \mu_\theta(Z)||_1] + W_f(p_\theta(x), p_D(x)) + \beta \cdot (||\theta||_1 + ||\phi||_1) - \lambda\, |\, \mathbb{E}_{Z\sim q_\phi(z)}[||Z - \mathbb{E}_{Z\sim q_\phi(z)}[Z]||_2^2] - v|, \quad (3)$$

where $\beta$ and $\lambda$ are hyperparamaters. Between each update step of (3) we also maximise (2) with respect to $f$.

For time-domain speech signals, similar correlations between adjacent samples are present throughout the signal. Consequently, the proposed model can use 1D-convolutional layers,

as the same kernel can be applied across the signals. This is ideal because convolutional layers have fewer free parameters than fully connected layers, thus preventing overfitting.

As mentioned, the proposed speech enhancement model will enhance blocks of audio. If the input and desired output block sizes are the same, then there is less information available about signal behaviour at the ends of the input block window than there is in the middle. Therefore, additional samples will be given as input, providing information about the signal behaviour at the window boundaries. N noisy samples will be used as input to enhance the central M.

# 4. Subjective Evaluation

Objective performance measures are often used to evaluate speech enhancement algorithms [6, 10]. However, such evaluations can be unreliable. For example, a recent paper found that while their low-rate speech coder performed worse according to POLQA [30], a subjective evaluation showed that their coder had similar performance to competing high-rate coders [31]. Hence we used a subjective test for the evaluation of our speech enhancement system. In this section, we first describe the dataset and model setup. Then, we discuss the subjective evaluation setup. Lastly, we present the listening test results.

## 4.1. Experimental Setup

In this section, we describe the experimental setup for SE-VCAE; this includes the dataset used, model architecture and the process for using the trained model to enhance speech.

### *4.1.1. Data Set*

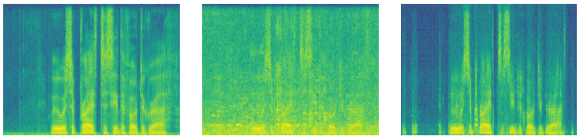
The dataset we used [32] consisted of 30 speakers taken from the Voice Bank [33] dataset. Of the 30 speakers, 28 were used to construct the training set, and two were used to build the testing set. All audio files were recorded at 42 kHz and were downsampled to 16 kHz. For both the training and testing sets, the average file length was three seconds. This dataset is identical to what was used by SE-GAN [6] and SE-WaveNet [10].

The noisy training data was corrupted using ten types of noise: two were artificially generated, and the remaining eight were taken from the Demand [34] dataset. Four different noise intensities were used, with an SNR of 15, 10, 5 and 0 dB. The noisy testing set was constructed using five types of noise from the Demand dataset, which were distinct from the eight that had already been used for the training data. Four different noise intensities were used, with SNRs of 17.5, 12.5, 7.5 and 2.5 dB.

### *4.1.2. Model Configuration*

In this section, we describe the implementation of SE-VCAE. The input and output block size of 62.5 and 37.5 milliseconds (i.e., $N = 1000$ and $M = 600$ samples) respectively. This setup allows SE-VCAE to outperform both SE-GAN and SE-WaveNet while utilising a shorter block size than the aforementioned models. Consequently, SE-VCAE is more feasible in a real-time situation. In this setup, the central 37.5-millisecond window of the input is the desired reconstruction.

During preliminary experiments, we found that VCAE trained on clean speech data had trouble reproducing high-frequency content. Applying a pre-emphasis filter (coefficient 0.95) to the input and target signals before splitting them into audio resolved this issue by amplifying the high-frequency information in the data (when testing, a de-emphasis filter is applied to the output). This technique was used in [6].



(a) *Clean.* (b) *Noisy.* (c) *Enhanced.*

Figure 2: *Spectrogram representations of the clean, noisy and enhanced signals for a single audio file. The signal had a length of* $\approx 2$ *seconds, and the SNR of the noisy signal was 2.5 dB.*

The encoder consisted of five 1D-convolutional layers with 32, 32, 64, 128, and 128 filters respectively. The first four layers used the Leaky-ReLU ($\alpha = 0.1$) activation, the last used a linear activation. Every layer had a kernel size of 31. The middle three layers had a stride of two, while the first and last had a stride of one. The output from the final convolutional layer was processed by a dense linear-layer with 330 output neurons, i.e., $z_{dim} = 330$. We define $\epsilon \sim \mathcal{N}(0, 0.05)$, and $v = 330$.

The decoder was given a tensor of latent feature vectors as input and first processed them with a dense linear-layer with $75 \cdot 128$ output neurons, the output of which was reshaped to be $(75, 128)$. Next, there were five transpose 1D-convolutional layers with 64, 32, 16, 16, and one filters, respectively. All the transpose convolutional layers had a kernel size of 31. The last layer used a linear activation, while the first four used the Leaky-ReLU ($\alpha = 0.1$) activation. The first three transpose convolutional layers had a stride of two, and the last two had a stride of one. Finally, the last transpose convolutional layer was followed by a dense linear-layer with 600 output neruons.

The function $f$ in the Wasserstein distance consisted of three 1D-convolutional layers, with 32, 64, and 128 filters respectively, and had batch-norm [35] intermediate layers. Each convolutional layer used a kernel size of 31, had a stride of two, and used the LeakyReLU ($\alpha = 0.1$) activation function. The output from the last convolutional layer was flattened and processed by a dense linear-layer with one output neuron.

The model was trained using Adam [36] with a learning rate of $1 \times 10^{-4}$, a batch size of 200, $\lambda = 0.01$, and $\beta = 1 \times 10^{-6}$.

### *4.1.3. Enhancing Testing Signals*

We now outline the procedure for using the trained SE-VCAE to enhance noisy signals. For a given audio file, we first applied a pre-emphasis filter (with coefficient 0.95) to the noisy test file. Then, we split the filtered signal into blocks of size $N$, such that the central $M$ samples of successive blocks overlap by $\frac{M}{2}$ samples. Applying SE-VCAE to this input yields enhanced blocks of size $M$, where successive enhanced blocks overlap by $\frac{M}{2}$ samples. Next, we joined the enhanced blocks together using a Hann window, smoothing any discontinuities between the blocks. Lastly, we applied a de-emphasis filter to the joined blocks to obtain the enhanced signal. Figure 2 gives an example of SE-VCAE enhancement quality. Three spectrograms are shown in this figure, one for each of the clean, noisy (2.5 dB SNR), and enhanced speech signals.

## 4.2. Evaluation Setup

In this section, we describe the setup for the subjective evaluation of SE-VCAE. First, the reference systems are introduced, and then we describe the subjective evaluation, for which we

Table 1: *Average scores (across all SNRs) with confidence intervals obtained from MUSHRA listening test.*

| Noisy | SE-GAN | SE-WaveNet | SE-VCAE |
|---|---|---|---|
| 26.9±3.2 | 50.1±3.1 | 48.0±3.7 | 59.0±3.4 |

used the Multiple Stimuli with Hidden Reference and Anchor (MUSHRA) [37] listing test, implemented using mushraJS [38].

#### *4.2.1. Reference Systems*

We use SE-GAN and SE-WaveNet as reference systems in our subjective evaluation. We chose these models because they are recent and also enhance time domain signals. Moreover, both systems were trained and tested on the same dataset as SE-VCAE (described in Section 4.1.1), meaning they can both be compared against our model in a single test. To obtain the comparison audio for SE-WaveNet, we used the samples available online,[1] and for SE-GAN, we used the pre-trained model (provided by the authors), available online.[2] For the implementation details of these models, see the respective papers.

#### *4.2.2. MUSHRA Listening Test Description*

The MUSHRA test consisted of 20 units, each unit presented the labelled clean audio file as a reference, and six other audio files with hidden labels. The six hidden audio files are: 1) The original noisy signal, 2) The enhanced noisy signal produced by SE-GAN, SE-WaveNet and SE-VCAE, 3) The clean audio file (this is the hidden reference), 4) The noisy speech signal but with an SNR 5dB less than the signal being enhanced [39] (this is the anchor). For each of the 20 units, the unlabelled audio files were in a randomised order. For each unit, the respondents were then asked to rate each of the hidden audio files with respect to the reference on a scale from 0 to 100.

### 4.3. Results

Next, we discuss the performance of the trained SE-VCAE. We had a total of six respondents for our MUSHRA listening test. Using these results, we investigated three situations. The first situation averages across all noisy signal SNRs, the second separates the results by SNR, and the third divides the results by noise type. To test for significant differences between the four models (noisy file, SE-GAN, SE-WaveNet and SE-VCAE), in all situations (i.e., overall and when split by noise type and SNR), we use a paired t-test with a p-value of $0.05$. We also display a 95% confidence interval around each mean score.

Table 1 shows the average score (across all SNRs and noise types) for the noisy signal and the enhanced signals from the three methods we are comparing. Based on the results from the paired t-tests, we can draw the following conclusions: 1) SE-GAN, SE-WaveNet and SE-VCAE all improve on the noisy signals, 2) SE-GAN and SE-WaveNet are statically equivalent, and 3) SE-VCAE improves upon both SE-GAN and SE-WaveNet.

Table 2 shows the mean scores and confidence intervals for each model, for each individual SNR. From these results and the statical significance tests we can draw the following conclusions: 1) SE-VCAE, SE-GAN and SE-WaveNet are preferred over the noisy signal for all SNRs, 2) SE-VCAE is preferred over SE-GAN and SE-WaveNet for SNR levels 2.5 dB and 7.5 dB, and 3) SE-VCAE is equivalent to SE-GAN and SE-WaveNet for SNR levels 12.5 dB and 17.5 dB respectively.

[1] http://jordipons.me/apps/speech-denoising-wavenet/

[2] https://github.com/santi-pdp/segan

Table 2: *Average scores with confidence intervals (split by SNR) obtained from MUSHRA listening test.*

| *SNR* (dB) | Noisy | SE-GAN | SE-WaveNet | SE-VCAE |
|---|---|---|---|---|
| 2.5 | 20.4±5.8 | 43.0±4.8 | 36.4±5.6 | 53.9±6.4 |
| 7.5 | 24.8±5.9 | 45.2±5.8 | 40.8±7.2 | 54.9±7.2 |
| 12.5 | 33.0±6.5 | 60.4±6.7 | 55.0±6.7 | 65.9±6.1 |
| 17.5 | 29.5±6.1 | 51.9±5.5 | 59.7±6.9 | 61.2±6.8 |

Table 3: *Average scores with confidence intervals (split by noise type) obtained from MUSHRA listening test.*

| *Noise* | Noisy | SE-GAN | SE-WaveNet | SE-VCAE |
|---|---|---|---|---|
| living | 24.0±7.0 | 44.2±5.8 | 36.1±7.1 | 63.5±6.9 |
| psquare | 25.2±5.1 | 44.8±4.2 | 46.6±5.7 | 54.7±5.6 |
| cafe | 29.8±5.7 | 60.7±5.7 | 54.0±6.3 | 61.6±6.3 |
| bus | 31.1±9.3 | 51.3±10.2 | 59.0±11.7 | 59.5±8.7 |

Lastly, we investigate how the noise type affects the performance of the enhancement methods. There are four different noise types in the 20 files used for the listening test, these are: inside a living room (living), public square (psquare), in a cafe (cafe), and inside a public bus (bus). The number of files and average SNR for each noise type is: 4 and 10.0 dB, 8 and 8.1 dB, 6 and 12.5 dB, 2 and 10.0 dB for living, psquare, cafe and bus respectively. Table 3 shows the mean scores and 95% confidence interval when the results are split by noise types. Based on the statistical significance testing, we can draw the following conclusions: 1) for all noise types, SE-VCAE, SE-GAN and SE-WaveNet improve on the noisy signal, 2) for living room and psquare, SE-VCAE improves on SE-GAN and SE-WaveNet, 3) for cafe, SE-VCAE and SE-GAN are statistically equivalent, but improve on SE-WaveNet, and 4) for bus, SE-VCAE, SE-GAN and SE-WaveNet are all statically equivalent.

The main results of this section are: 1) Overall, SE-VCAE improves on SE-GAN, SE-WaveNet and the original noisy files, 2) SE-VCAE outperforms SE-GAN and SE-WaveNet for SNR levels of 2.5dB and 7.5dB, 3) all three enhancement methods are preferred to even the high (17.5dB) SNR noisy files, and 4) for all noise types, SE-VCAE is either equivalent to or better than the other enhancement methods.

## 5. Conclusion

In conclusion, we proposed the Speech Enhancement Variance Constrained Autoencoder (SE-VCAE), a (time-domain) speech enhancement system, based on the Variance Constrained Autoencoder (VCAE). We demonstrated using a subjective test that our proposed SE-VCAE outperforms SE-GAN and SE-WaveNet, two recent speech enhancement systems. Moreover, our proposed model has a simple neural network structure compared to these competing methods. Our results show that it is possible to enhance noisy speech without overly complex neural network structures.

## 6. Acknowledgements

This work was supported by funding from GN.

# 7. References


[1] P. C. Loizou, *Speech enhancement: theory and practice*. CRC press, 2007.

[2] F. Weninger, H. Erdogan, S. Watanabe, E. Vincent, J. Le Roux, J. R. Hershey, and B. Schuller, "Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR," in *Proceedings of the International Conference on Latent Variable Analysis and Signal Separation*. Springer, 2015, pp. 91–99.

[3] Y. Xu, J. Du, L. Dai, and C. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 23, no. 1, pp. 7–19, 2015.

[4] A. Kumar and D. Florencio, "Speech enhancement in multiple-noise conditions using deep neural networks," in *Interspeech 2016*, 2016, pp. 3738–3742.

[5] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder," in *Proc. Interspeech 2013*, 2013, pp. 436–440.

[6] S. Pascual, A. Bonafonte, and J. Serra, "SEGAN: Speech enhancement generative adversarial network," in *Proc. Interspeech 2017*, 2017, pp. 3642–3646.

[7] S. R. Park and J. Lee, "A fully convolutional neural network for speech enhancement," in *Proc. Interspeech 2017*, 2017, pp. 1993–1997.

[8] H. Zhao, S. Zarar, I. Tashev, and C. Lee, "Convolutional-recurrent neural networks for speech enhancement," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 2401–2405.

[9] Y. Bando, M. Mimura, K. Itoyama, K. Yoshii, and T. Kawahara, "Statistical speech enhancement based on probabilistic integration of variational autoencoder and non-negative matrix factorization," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 716–720.

[10] D. Rethage, J. Pons, and X. Serra, "A WaveNet for speech denoising," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5069–5073.

[11] Y. Luo and N. Mesgarani, "TasNet: time-domain audio separation network for real-time, single-channel speech separation," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 696–700.

[12] Y. Luo and N. Mesgarani, "TasNet: Surpassing ideal time-frequency masking for speech separation," *arXiv preprint arXiv:1809.07454*, 2018.

[13] D. Griffin and J. Lim, "Signal estimation from modified short-time Fourier transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 2, pp. 236–243, 1984.

[14] J. Lim and A. Oppenheim, "All-pole modeling of degraded speech," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 26, no. 3, pp. 197–210, 1978.

[15] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, 2014, pp. 2672–2680.

[16] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "WaveNet: A generative model for raw audio," in *9th ISCA Speech Synthesis Workshop*, 2016, pp. 125–125.

[17] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017, pp. 214–223.

[18] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *Proceedings of the 31st International Conference on Machine Learning (ICML)*, 2014.

[19] D. J. Rezende, S. Mohamed, and D. Wierstra, "Stochastic backpropagation and approximate inference in deep generative models," in *Proceedings of the 31st International Conference on Machine Learning (ICML)*, 2014, pp. 1278–1286.

[20] L. Pandey, A. Kumar, and V. Namboodiri, "Monoaural audio source separation using variational autoencoders," *Proc. Interspeech 2018*, pp. 3489–3493, 2018.

[21] S. Leglaive, L. Girin, and R. Horaud, "A variance modelling framework based on variational autoencoders for speech enhancement," in *2018 IEEE 28th International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE, 2018, pp. 1–6.

[22] S. Zhao, J. Song, and S. Ermon, "InfoVAE: Balancing learning and inference in variational autoencoders," *arXiv preprint arXiv:1706.02262v3*, 2018.

[23] A. A. Alemi, B. Poole, I. Fischer, J. V. Dillon, R. A. Saurous, and K. Murphy, "Fixing a broken ELBO," in *Proceedings of the 35th International Conference on Machine Learning (ICML)*, 2018, pp. 159–168.

[24] D. T. Braithwaite, M. O'Connor, and W. B. Kleijn, "Variance constrained autoencoding," *To Appear*, 2019.

[25] I. Tolstikhin, O. Bousquet, S. Gelly, and B. Schoelkopf, "Wasserstein auto-encoders," in *Proceedings of the 6th International Conference on Learning Representations (ICLR)*, 2018.

[26] D. T. Braithwaite and W. B. Kleijn, "Bounded information rate variational autoencoders," *arXiv preprint arXiv:1807.07306*, 2018.

[27] C. Villani, *Optimal transport: old and new*. Springer Science & Business Media, 2008, vol. 338.

[28] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of wasserstein GANs," in *Proceedings of Advances in Neural Information Processing Systems*, 2017, pp. 5767–5777.

[29] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.

[30] International Telecommunications Union, "Rec. p.863: Perceptual objective listening quality prediction," Tech. Rep.

[31] W. B. Kleijn, F. S. C. Lim, A. Luebs, J. Skoglund, F. Stimberg, Q. Wang, and T. C. Walters, "WaveNet based low rate speech coding," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 676–680.

[32] C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi, "Investigating rnn-based speech enhancement methods for noise-robust text-to-speech," in *Proceedings of the 9th ISCA Speech Synthesis Workshop*, 2016, pp. 146–152.

[33] C. Veaux, J. Yamagishi, and S. King, "The voice bank corpus: Design, collection and data analysis of a large regional accent speech database," in *Oriental COCOSDA held jointly with 2013 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE), 2013 International Conference*. IEEE, 2013, pp. 1–4.

[34] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multi-channel acoustic noise database: A database of multichannel environmental noise recordings," *The Journal of the Acoustical Society of America*, vol. 133, no. 5, pp. 3591–3591, 2013.

[35] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proceedings of the 32nd International Conference on Machine Learning*, 2015, pp. 448–456.

[36] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014.

[37] International Telecommunications Union, "Method for the subjective assessment of intermediate quality levels of coding systems," Tech. Rep.

[38] C. Baume. (2012) mushraJS. [Online]. Available: https://github.com/chrisbaume/mushraJS

[39] F. Deng and C. Bao, "Speech enhancement based on bayesian decision and spectral amplitude estimation," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2015, no. 1, p. 28, 2015.