Speech Enhancement with Variance Constrained Autoencoders

Code: https://github.com/danielbraithwt/Speech-Enhancement-with-Variance-Constrained-Autoencoders

Daniel T. Braithwaite and W. Bastiaan Kleijn
School of Engineering and Computer Science, Victoria University of Wellington, New Zealand

Problem

- Magnitude spectrum enhancement methods provide spectrogram output:
  - Quality of enhancement relies on method used to synthesise speech from spectrogram.
- Classical enhancement paradigm (e.g., Wiener filtering):
  - Methods do not know attributes of speech.
  - Methods produce audible distortions.
- Consider a situation where multiple speech signals are equally plausible (given some noisy input):
  - MMSE approach compromises between them, which yields distortions.
- Instead, generative enhancement paradigm:
  - Generate convincing speech that matches the content of the original noisy signal.
  - In the previous example a generative approach picks one of the equally likely signals, leading to reasonable sounding speech.
  - A Maximum likelihood approach would also do this, but without a prior on clean speech the noisy signal would be returned.
  - Generative approach includes a prior.
- Other time domain enhancement approaches based off generative models are not truly generative.
- Objective: Develop a time domain speech enhancement model based on the Variance Constrained Autoencoder (VCAE) which is a step towards a generative system and is computationally less complex than competing solutions.

Contributions

- The Speech Enhancement Variance Constrained Autoencoder (SE-VCAE).
- SE-VCAE outperforms SE-GAN and SE-WaveNet
  - A subjective MUSHRA evaluation demonstrates that SE-VCAE is better at de-noising speech, compared to SE-GAN and SE-WaveNet.
  - SE-VCAE uses a less complex neural network architecture compared to SE-GAN and SE-WaveNet.

Background

- Speech Enhancement GAN (SE-GAN) [1]:
  - Based on the Generative Adversarial Network (GAN).
  - Adds an encoder and L1 error criterion to ensure that generated audio matches desired clean speech.
  - Complex network structure.
- Speech Enhancement WaveNet (SE-WaveNet) [2]:
  - Based on WaveNet.
  - Minimises regression loss function, not generative.
  - Complex network structure.
- Both SE-GAN and SE-WaveNet outperform the classical Wiener enhancement method, according to subjective test.

Variance Constrained Autoencoder

$$\varepsilon \sim p_\phi(\varepsilon)$$

$$x \sim p_D(x) \rightarrow \boxed{\mu_\phi(x)} \rightarrow \oplus \rightarrow z \rightarrow \boxed{p_\theta(x|z)} \rightarrow \hat{x}$$

- $q_\phi(z|x)$ is the distribution over latent variables for a given x, sampling from $q_\phi(z|x)$ is defined as:
  - $\mu_\phi(x)$ outputs the mean of $q_\phi(z|x)$, μ_ϕ is implemented by a neural network with parameter ϕ.
  - sample $\varepsilon \sim p(\varepsilon)$, where $p(\varepsilon)$ is a user-defined distribution (e.g., $\mathcal{N}(0, I \cdot \sigma_\varepsilon)$).
  - then, $z \sim q_\phi(z|x)$ has the form $z = \mu_\phi(x) + \varepsilon$.
- $p_\theta(x|z)$ is the distribution over data given the latent features, typically deterministic.
  - Typically, $p_\theta(X|Z)$ is assumed to be a factored Gaussian or Laplace distribution.
- $q_\phi(z|x)$ and $p_\theta(x|z)$ denoted encoder and decoder.
- Optimise by maximising likelihood of data, subject to a variance constraint:
  - Cannot constrain variance directly, need to use a penalty function.

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

Speech Enhancement VCAE

- Objective: Use VCAE for speech enhancement.
- Define the input distribution as $\tilde{X}$, the noisy data.
- Output distribution is the clean data.
- The objective function is an extension of VCAEs:
  - Maximises likelihood of clean data at output given features inferred from corresponding noisy data.
  - $p_\theta(X|Z)$ assumed to be factored Laplace distribution.
  - Constrains the latent distributions variance using a penalty function.
  - Applies L1 regularisation to the weights of the encoder and decoder.
  - Minimises the Wasserstein distance between $p_\theta(x)$ and $p_D(x)$ (given by $W_f(p_\theta(x), p_D(x))$).

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

Dataset

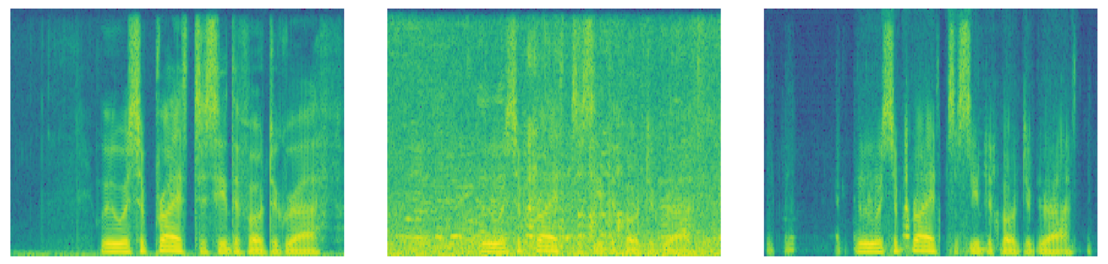
- The dataset consisted of 30 speakers from VoiceBank [3]:
  - 28 for the training set, and two for the testing set.
- Noisy data constructed by corrupting the training and testing speakers with noise; some noise is artificially generated and the remainder is from the Demand dataset [4]:
  - Training and testing noise types are distinct.
  - Training noise SNRs are 15, 10, 5, and 0 dB.
  - Testing noise SNRs are 17.5, 12.5, 7.5, and 2.5 dB.
- Audio recorded at 42 kHz, and down-sampled to 16 kHz.
- Same dataset as was used by SE-GAN and SE-WaveNet.

Experimental Set-up

- Output block size was 37.5 milliseconds (600 samples).
- Input block size was 62.5 milliseconds (1000 samples). Central 37.5 milliseconds is the desired reconstruction.
- Before splitting audio files into blocks, a pre-emphasis (0.95) filter is applied.
- The encoder/decoder neural networks used:
  - 1D-Convolutional layers; Batch norm; 330 latent features.
- The procedure for enhancing a test signal is:
  - Apply a pre-emphasis (0.95) filter to the audio file.
  - Split the file into blocks of length 1000, such that for successive blocks the central 600 samples overlap by 300 samples.
  - Apply the trained SE-VCAE model to these blocks, yielding the set of enhanced blocks.
  - Join the enhanced blocks using a Hann window (this smooths any discontinuities between blocks).
  - Apply de-emphasis (0.95) filter to the joined blocks.

Example of Enhancement

- Spectrogram representations of the clean, noisy and SE-VCAE enhanced versions of a single audio file.
- Length is ≈ 2 seconds. SNR is 2.5 dB.



Clean. Noisy. Enhanced.

MUSHRA Set-up

- Subjective MUSHRA test used to evaluate model.
- SE-GAN and SE-WaveNet were reference systems.
- 20 units in total. One unit consists of six hidden audio files:
  - noisy signal; enhancement produced by SE-VCAE/GAN/WaveNet; hidden reference; noisy speech signal with 5 dB lower SNR.
- Six respondents.
- Used a paired t-test for significance testing, with a p-value of 0.05.

Overall Results

- Average MUSHRA scores for each model tested, over all noise types and SNRs.

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

- All models outperform the noisy signals.
- SE-GAN and SE-WaveNet are equivalent.
- SE-VCAE improves upon SE-GAN and SE-WaveNet.

Results per SNR

- Average MUSHRA scores for each model tested, split by SNR.

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**

- All models outperform the noisy signals for all SNRs.
- SE-VCAE outperforms SE-GAN and SE-WaveNet for the SNRs 2.5 dB and 7.5 dB.
- SE-VCAE is equivalent to SE-GAN and SE-WaveNet for the SNRs 12.5 dB and 17.5 dB.

Results per Noise Type

- Average MUSHRA scores for each model tested, split by noise type.

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**

- All models outperform the noisy signals for all noise types.
- For living room and psquare, SE-VCAE outperforms SE-GAN and SE-WaveNet.
- For cafe SE-VCAE is equivalent to SE-GAN, both improve on SE-WaveNet.
- For bus, all models are equivalent.

Conclusions

- SE-VCAE outperforms both SE-GAN and SE-WaveNet.
  - Overall SE-VCAE improves upon SE-GAN and SE-WaveNet, shown by subjective evaluation.
  - When scores are split by SNR: SE-VCAE improves upon SE-GAN and SE-WaveNet for SNRs of 2.5 dB and 7.5 dB.
- SE-VCAE uses a less complex neural network structure than these two competing models.
  - SE-VCAE has encoder and decoder networks that consist of fewer layers.

Acknowledgements

This work was supported by funding from GN.

References

[1] S. Pascual, A. Bonafonte, and J. Serra, "SEGAN: Speech enhancement generative adversarial network," in *Proc. Interspeech 2017*, 2017, pp. 3642–3646.
[2] 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.
[3] 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.
[4] 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.