*INTERSPEECH 2020*
October 25–29, 2020, Shanghai, China

# Speech Enhancement with Stochastic Temporal Convolutional Networks

*Julius Richter, Guillaume Carbajal, Timo Gerkmann*

Signal Processing (SP), Universität Hamburg, Germany
{julius.richter, guillaume.carbajal, timo.gerkmann}@uni-hamburg.de

## Abstract

We consider the problem of speech modeling in speech enhancement. Recently, deep generative approaches based on variational autoencoders have been proposed to model speech spectrograms. However, these approaches are based either on hierarchical or temporal dependencies of stochastic latent variables. In this paper, we propose a generative approach to speech enhancement based on a stochastic temporal convolutional network, which combines both hierarchical and temporal dependencies of stochastic variables. We evaluate our method with real recordings of different noisy environments. The proposed speech enhancement method outperforms a previous non-sequential approach based on feed-forward fully-connected networks in terms of speech distortion, instrumental speech quality and intelligibility. At the same time, the computational cost of the proposed generative speech model remains feasible, due to inherent parallelism of the convolutional architecture.

**Index Terms**: speech enhancement, stochastic temporal convolutional networks, generative model, variational inference.

## 1. Introduction

Speech processing systems such as mobile phones, VoIP, teleconferencing systems, speech recognition, and hearing aids require improving speech quality and intelligibility [1]. Single-channel speech enhancement aims at recovering a clean speech signal from a mixture which can contain speech and additive noise [2]. In traditional speech enhancement, Bayesian estimators are often used to estimate speech coefficients [3], and speech parameters such as the speech power spectral density [4], the noise power spectral density [5], and the phase [6].

Besides Bayesian estimators, a Bayesian generative model of the observed signal can be built based on available knowledge about its production process [7]. The resulting probabilistic speech enhancement methods aim to mimic the hidden random process of speech and may be used to generate artificial data that resembles the properties of a given dataset.

Several speech enhancement methods combine concepts of Bayesian inference and deep learning, e.g. [8–10]. These approaches make use of generative models, in particular the variational autoencoder (VAE) [11, 12]. Similar to methods based on denoising autoencoders [13, 14], VAE-based methods are also capable of denoising speech spectrograms by using stochastic latent variables, which are hierarchically arranged in a top-down fashion. Although both achieve good performance compared against traditional speech enhancement methods, most of them only reconstruct single spectrogram frames without modeling temporal dependencies.

This work has been funded by the German Research Foundation (DFG) in the transregio project Crossmodal Learning (TRR 169) and ahoi.digital. We would like to thank Rohde & Schwarz SwissQual AG for their support with POLQA.

Recently, a generative model called stochastic temporal convolutional network (STCN) was proposed which can simultaneously capture temporal dependencies of variable-length sequences and learn correlations between output variables [15]. The STCN uses a top-down hierarchy of stochastic latent variables which are conditioned on deterministic representations computed bottom-up. The deterministic representations correspond to dilated convolutions of a temporal convolutional network (TCN) [16]. Lateral shortcut connections between the deterministic and latent variables allow higher levels of the latent hierarchy to focus on more abstract invariant features [17]. The network architecture is capable of generating high-quality synthetic samples and achieves state-of-the-art log-likelihoods in speech synthesis. However, to our best knowledge, STCNs have not been applied in speech enhancement yet.

In this work, motivated by the aforementioned advantages, we propose a generative approach to speech enhancement using an STCN as a speaker-independent speech model to estimate the variance of clean speech. The estimation of the noise variance, on the other hand, is based on a non-negative matrix factorization (NMF) [18, 19]. We compare the proposed convolutional approach to a non-sequential approach based on a feed-forward fully-connected architecture using the same NMF parameter optimization algorithm [10].

The rest of this paper is organized as follows: in Section 2 we introduce a general speech enhancement method based on speech and noise variance estimation. Section 3 describes deep generative speech models based on variational inference. The novel STCN architecture is introduced in Section 4 followed by the evaluation in Section 5.

## 2. Speech Enhancement

The general goal of single-channel speech enhancement is defined as recovering a speech signal from an observed signal involving interfering sound sources or reverberation. For this purpose, a common approach consists in estimating the speech and noise variances in order to reconstruct the clean speech signal.

### 2.1. Model

In the time-frequency domain using the short time Fourier transform (STFT), the mixture signal $y_{ft} \in \mathbb{C}$ is the sum of the clean speech $s_{ft} \in \mathbb{C}$ and the noise $n_{ft} \in \mathbb{C}$, such that

$$y_{ft} = s_{ft} + n_{ft}, \tag{1}$$

for every frequency bin $f \in \{1, \dots, F\}$ and time frame $t \in \{1, \dots, T\}$, where $F$ denotes the number of frequency bins and $T$ the number of time frames of the utterance. The signals $s_{ft}$ and $n_{ft}$ are modeled as mutually independent circularly-symmetric complex Gaussian random variables,

$$s_{ft} \sim \mathcal{CN}(0, \sigma^2_{s,ft}), \quad n_{ft} \sim \mathcal{CN}(0, \sigma^2_{n,ft}), \tag{2}$$



http://dx.doi.org/10.21437/Interspeech.2020-2588

where $\mathcal{CN}(\mu, \sigma^2)$ denotes a complex Gaussian distribution with mean $\mu$ and variance $\sigma^2$. Under a local stationary assumption, the variances $\sigma^2_{s,ft}$ and $\sigma^2_{n,ft}$ represent the short-time power spectral density of $s_{ft}$ and $n_{ft}$, respectively [20]. Given the noisy mixture, clean speech coefficients can be estimated in the minimum mean square sense using the Wiener estimator

$$\hat{s}_{ft} = \frac{\hat{\sigma}^2_{s,ft}}{\hat{\sigma}^2_{s,ft} + \hat{\sigma}^2_{n,ft}} \, y_{ft} \tag{3}$$

where $\hat{\sigma}^2_{s,ft}$ and $\hat{\sigma}^2_{n,ft}$ are the estimated variances using the corresponding signal variance models as described next.

### 2.2. Signal variance models

NMF is a popular choice for modeling signal variances based on previously trained clean speech spectra [18]. However, it is limited in its modeling capacity due to the linear parametrization of the variances. Therefore, we use a non-linear deep generative model for speech variance estimation which depends on stochastic latent variables. The noise variance, on the other hand, is estimated based on an untrained NMF noise model.

### 2.3. Robust mixture model

In order to provide some robustness with respect to the time-varying loudness of different speech signals, the mixture model in Eq. (1) is extended with a frequency-independent but time-varying gain $g_t \in \mathbb{R}_+$ [10], such that

$$y_{ft} = \sqrt{g_t}\, s_{ft} + n_{ft}. \tag{4}$$

## 3. Generative Speech Models

### 3.1. Variational Autoencoders

With the original VAE framework [11, 12], speech power coefficients $x_{ft} = |s_{ft}|^2$ are created by a random process, involving an unobserved random variable $\mathbf{z}_t \in \mathbb{R}^D$. This process is shown in Fig. 1a and consists of two steps: first a value $\mathbf{z}_t^{(i)}$ is drawn from a prior probability distribution $p(\mathbf{z}_t)$ as the $i$-th sample, and second a power spectrogram frame $\mathbf{x}_t^{(i)} \in \mathbb{R}_+^F$ is generated from a conditional probability distribution $p_\theta(\mathbf{x}_t|\mathbf{z}_t)$ which is also called generative distribution.

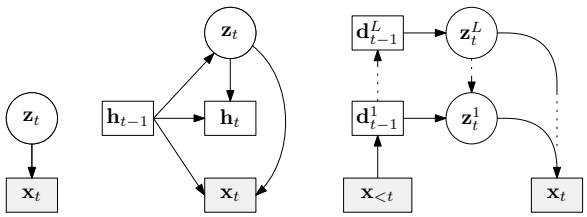
In variational inference, a recognition model $q_\phi(\mathbf{z}_t|\mathbf{x}_t)$ is introduced as an approximation to the intractable true posterior $p(\mathbf{z}_t|\mathbf{x}_t)$ [21]. All distributions are modeled as neural networks and its parameters $\theta$ and $\phi$ are jointly optimized maximizing the variational lower bound on the marginal log-likelihood for a given spectrogram frame

$$\begin{aligned}\log p(\mathbf{x}_t^{(i)}) \geq &- D_{\mathrm{KL}}(q_\phi(\mathbf{z}_t|\mathbf{x}_t^{(i)}) \,||\, p(\mathbf{z}_t)) \\ &+ \mathbb{E}_{q_\phi(\mathbf{z}_t|\mathbf{x}_t^{(i)})}[\log p_\theta(\mathbf{x}_t^{(i)}|\mathbf{z}_t)],\end{aligned} \tag{5}$$

where $D_{\mathrm{KL}}$ denotes the Kullback-Leibler divergence between the approximate posterior $q_\phi(\mathbf{z}_t|\mathbf{x}_t^{(i)})$ and the prior probability distribution $p(\mathbf{z}_t)$ which is commonly chosen to be a standard Gaussian distribution with zero mean and unit variance. The second term in Eq. (5) is an expected reconstruction error and requires estimation by sampling.

The statistical properties of clean speech, which are learned by the generative model during training, can be used to estimate the distribution of the noisy mixture coefficients $y_{ft}$. Since speech and noise signals are supposed to be mutually independent given the latent variable $\mathbf{z}_t$, we have

$$y_{ft}|\mathbf{z}_t \sim \mathcal{CN}(0, g_t \hat{\sigma}^2_{s,ft}(\mathbf{z}_t) + \hat{\sigma}^2_{n,ft}). \tag{6}$$




(a) *VAE* (b) *VRNN* (c) *STCN*

Figure 1: *Generative models: (a) the variational autoencoder (VAE), (b) the variational recurrent neural network (VRNN), and (c) the stochastic temporal convolutional network (STCN). Circles are stochastic variables, squares are deterministic variables, and gray boxes represent the observed sequence.*

Straightforward maximum likelihood estimation of Eq. (6) is intractable due to the non-linear relation between the speech variance and the latent variables. However, a Monte Carlo expectation maximization (MCEM) algorithm can be employed which iteratively optimizes the NMF noise parameters and the time-varying gain $g_t$ using a block-coordinate approach [10]. Samples of the latent variable are drawn from the approximate posterior distribution using the random walk Metropolis-Hastings algorithm [22].

### 3.2. Temporal dependencies

In the VAE framework [11, 12], the generative model does not consider temporal dependencies of sequential data. Thus, we introduce the variational recurrent neural network (VRNN) as a common deep generative model capable of modeling the temporal dependencies by introducing a deterministic hidden state $\mathbf{h}_t \in \mathbb{R}^H$ which is updated in a recurrent fashion [23]. The generation of new samples and the recurrence of the hidden state is illustrated in Fig. 1b.

VRNNs have an infinite internal memory due to occurring feedback loops. As we are interested in modeling finite length phonemes, this infinite internal memory may be seen as a conceptual disadvantage. Furthermore, due to the necessary back-propagation through time, recurrent network architectures are rather slow to train and may also suffer from the vanishing gradient problem [24].

## 4. Stochastic Temporal Convolutional Networks

The STCN is another generative model capable of modeling the joint probability distribution of variable-length sequences. For this purpose, the architecture employs two main modules: 1) a TCN with deterministic representations computed bottom-up; and 2) a stochastic latent variable hierarchy with top-down dependencies, as illustrated in Fig. 1c.

### 4.1. Temporal convolutional network

In the TCN, dilated causal convolutions are applied over the input sequence $\mathbf{x}$ to compute a set of deterministic representations $\mathbf{d} = \{\mathbf{d}^1, \ldots, \mathbf{d}^L\}$ where $L$ is the total number of stacked layers with corresponding sequences $\mathbf{d}^l = (\mathbf{d}^l_1, \ldots, \mathbf{d}^l_T)$ and $\mathbf{d}^l_t \in \mathbb{R}^{D_l}$. The computation of the deterministic representations is bottom-up and recursively defined as $\mathbf{d}^l_t = f(\mathbf{d}^{l-1}_t)$ starting with $\mathbf{d}^0 = \mathbf{x}$. The function $f$ is a series of transformations containing residual connections, dilated convolutions, weight normalization [25], ReLUs, and spatial dropout [26].

Using larger dilation factors in the convolution enables a deterministic representation at higher levels to represent a wider range of inputs, thus effectively expanding the receptive field compared to a standard convolution. Therefore, contrary to VRNNs, the STCN has the advantage to accurately set a desired receptive field size, which may be useful for the task of speech enhancement. The overall receptive field size is given as $(k-1)(2^L-1)$ with filter size $k$ and can be set in line with the typical length of phonemes, or in line with the temporal integration time of the human auditory system, which has an upper bound of a few hundreds of milliseconds [27].

Furthermore, residual and skip connections allow local information to propagate through the network while avoiding the vanishing gradient problem [28, 29].

### 4.2. Stochastic latent variable hierarchy

Stochastic latent variables are arranged in correspondence to the TCN layers. Thus, there also exists a set of random variables $\mathbf{z} = \{\mathbf{z}^1, \dots, \mathbf{z}^L\}$ with sequences $\mathbf{z}^l = (\mathbf{z}_1^l, \dots, \mathbf{z}_T^l)$ and $\mathbf{z}_t^l \in \mathbb{R}^{Z_l}$ which capture temporal dependencies at different time scales. The decoupling of deterministic and stochastic layers is shown in Fig. 2. The stochastic latent variable hierarchy can be seen as a modular add-on for any temporal convolutional network architecture.

### 4.3. Inference

The recognition model of the STCN relies on a top-down dependency of the latent variables, as illustrated in Fig. 2. As a result, the parameters of the approximate posterior for each latent layer $l$ are computed by

$$[\hat{\boldsymbol{\mu}}_{t,q}^l, \hat{\mathbf{v}}_{t,q}^l] = \begin{cases} f_q^{(l)}(\mathbf{z}_t^{l+1}, \mathbf{d}_t^l), & \text{for } l \in [1, L-1] \\ f_q^{(L)}(\mathbf{d}_t^L), & \text{for } l = L, \end{cases} \tag{7}$$

where $\{f_q^{(l)}\}_{l=1,\dots,L}$ is a set of neural networks consisting of stacked layers of 1D convolutions with kernel size 1.

This is contrary to ordinary VAEs, where the inference is defined as a bottom-up process. Furthermore, the mean $\hat{\boldsymbol{\mu}}_{t,q}^l$ and the diagonal variance entries $\hat{\mathbf{v}}_{t,q}^l$ are subsequently corrected by precision-weighted addition [17], such that

$$\begin{aligned} \boldsymbol{\mu}_{t,q}^l &= \mathbf{v}_{t,q}^l(\hat{\boldsymbol{\mu}}_{t,q}^l(\hat{\mathbf{v}}_{t,q}^l)^{-2} + \boldsymbol{\mu}_{t,p}^l(\mathbf{v}_{t,p}^l)^{-2}), \\ \mathbf{v}_{t,q}^l &= \frac{1}{(\hat{\mathbf{v}}_{t,q}^l)^{-2} + (\mathbf{v}_{t,p}^l)^{-2}}, \end{aligned} \tag{8}$$

where $\boldsymbol{\mu}_{t,p}^l$ and $\mathbf{v}_{t,p}^l$ are the parameters of the prior distribution which is described in Sec. 4.4.

Finally, the approximate posterior distribution of the set of latent variables $\mathbf{z}_t = \{\mathbf{z}_t^1, \dots, \mathbf{z}_t^L\}$ conditioned on the set of deterministic representations $\mathbf{d}_t = \{\mathbf{d}_t^1, \dots, \mathbf{d}_t^L\}$ at time step $t$ is given as

$$q(\mathbf{z}_t|\mathbf{d}_t) = q(\mathbf{z}_t^L|\mathbf{d}_t^L) \prod_{l=1}^{L-1} q(\mathbf{z}_t^l|\mathbf{z}_t^{l+1}, \mathbf{d}_t^l), \tag{9}$$

where each distribution of every latent layer $l$ is modeled as a multidimensional Gaussian $\mathcal{N}(\boldsymbol{\mu}_{t,q}^l, \mathrm{diag}(\mathbf{v}_{t,q}^l))$.

### 4.4. Prior distribution

The prior distribution of the set of latent variables $\mathbf{z}_t$ at time step $t$ depends on the set of deterministic representations $\mathbf{d}_{t-1}$ of the previous time step and is layer-wise factorized as

$$p(\mathbf{z}_t|\mathbf{d}_{t-1}) = p(\mathbf{z}_t^L|\mathbf{d}_{t-1}^L) \prod_{l=1}^{L-1} p(\mathbf{z}_t^l|\mathbf{z}_t^{l+1}, \mathbf{d}_{t-1}^l), \tag{10}$$

where each distribution of every latent layer $l$ is modeled as a multidimensional real-valued Gaussian $\mathcal{N}(\boldsymbol{\mu}_{t,p}^l, \mathrm{diag}(\mathbf{v}_{t,p}^l))$ with mean and diagonal variance entries

$$[\boldsymbol{\mu}_{t,p}^l, \mathbf{v}_{t,p}^l] = \begin{cases} f_p^{(l)}(\mathbf{z}_t^{l+1}, \mathbf{d}_{t-1}^l), & \text{for } l \in [1, L-1] \\ f_p^{(L)}(\mathbf{d}_{t-1}^L), & \text{for } l = L, \end{cases} \tag{11}$$

and the set of neural networks $\{f_p^{(l)}\}_{l=1,\dots,L}$ consisting of stacked layers of 1D convolutions with kernel size 1.

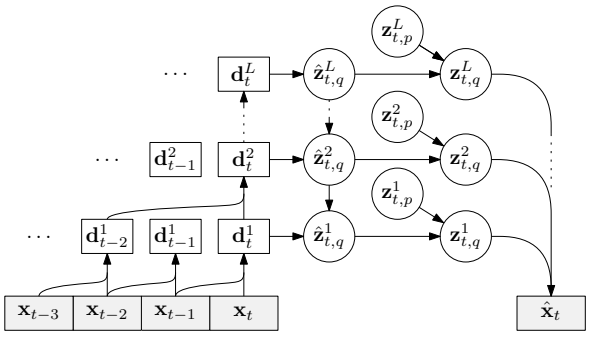



Figure 2: *Computational graph of the inference based on the prior and the approximate posterior distribution.*

### 4.5. Observation model

The final prediction of the clean speech coefficients $\hat{s}_{ft}$ given the latent variable $\mathbf{z}_t$ is modeled as

$$\hat{s}_{ft}|\mathbf{z}_t \sim \mathcal{CN}(0, \{\hat{\boldsymbol{\sigma}}_{s,t}^2(\mathbf{z}_t)\}_f)) \tag{12}$$

where $\hat{\boldsymbol{\sigma}}_{s,ft}^2 : \mathbb{R}^D \mapsto \mathbb{R}_+^F$ represents a multidimensional nonlinear function modeled by stacked TCN layers. In order to take all latent variables into account, we use the concatenation of samples from all latent layers $\mathbf{z}_t = (\mathbf{z}_t^1 \cdots \mathbf{z}_t^L)$, which takes inspiration from recent convolutional architectures [30].

### 4.6. Learning objective

The learning objective of the STCN is similar to Eq. (5), namely optimizing the variational lower bound on the log-likelihood at time step $t$, resulting in the loss function

$$\mathcal{L}(\mathbf{x}_t) = \mathcal{L}_t^{\mathrm{KL}}(\mathbf{x}_t) + \mathcal{L}_t^{\mathrm{Recon}}(\mathbf{x}_t). \tag{13}$$

Using the factorizations from Eq. (9) and (10), the regularization term becomes

$$\begin{aligned} \mathcal{L}_t^{\mathrm{KL}}(\mathbf{x}_t) = \; & D_{\mathrm{KL}}(q(\mathbf{z}_t^L|\mathbf{d}_t^L) \,||\, p(\mathbf{z}_t^L|\mathbf{d}_{t-1}^L)) \\ & + \sum_{l=1}^{L-1} \mathbb{E}_{q(\mathbf{z}_t^{l+1}|\cdot)}[D_{\mathrm{KL}}(q(\mathbf{z}_t^l|\mathbf{z}_t^{l+1}, \mathbf{d}_t^l) \,||\, p(\mathbf{z}_t^l|\mathbf{z}_t^{l+1}, \mathbf{d}_{t-1}^l))], \end{aligned} \tag{14}$$

where the expectation over the latent variables $\mathbf{z}_t^{l+1}$ is approximated using the reparameterization trick [11]. The Kullback-Leibler divergences can be calculated analytically since all appearing distributions are diagonal multivariate Gaussians.

The output of $\hat{\boldsymbol{\sigma}}^2_{s,t}(\mathbf{z}_t)$ also represents the estimated short-time power spectral density $\hat{\mathbf{x}}_t$ at frame $t$, therefore the reconstruction loss is defined as

$$\mathcal{L}_t^{\text{Recon}}(\mathbf{x}_t) = \mathbb{E}_{q(\mathbf{z}_t|\mathbf{x}_{\leq t})}\left[\left(\log \frac{\mathbf{x}_t + \epsilon}{\hat{\boldsymbol{\sigma}}^2_{s,t}(\mathbf{z}_t) + \epsilon}\right)^2\right] \quad (15)$$

where the expectation is approximated using the reparameterization trick and $\epsilon$ is numerically motivated in order not divide by zero.

# 5. Evaluation

In this section, we compare the performance of the proposed STCN approach with a non-sequential VAE approach [10]. We assess the performance in terms of speech distortion, speech intelligibility, and speech quality.

Very recently Leglaive et al. proposed a speech enhancement method based on VRNNs [31]. However, their approach is based on a different optimization algorithm than in the non-sequential VAE and the STCN considered here, which would make a direct comparison of the approaches difficult.

We also refrain from comparing with the other popular generative speech model which is based on generative adversarial networks [8], because its training needs clean and noisy speech pairs. Thus, the model is restricted to a limited set of noise types which is not the case in our approach. Furthermore, the number of learnable parameters differs by two orders of magnitude compared to the STCN approach.

## 5.1. Dataset

As training data we use approximately 25 hours of clean speech from the "si_tr_s" subset of the Wall Street Journal (WSJ0) dataset [32]. For testing we use 651 synthetic mixtures corresponding to approximately 1.5 hours of noisy speech. The clean speech signals are taken from the "si_et_05" subset of WSJ0 (8 unseen speakers), and the noise signals from the "verification" subset of the QUT-NOISE dataset [33]. Each mixture is created by uniformly sampling a noise type among "cafe", "home", "street", "car" and a signal-to-noise ratio (SNR) among -5, 0, 5 dB. All signals have a sampling rate of 16 kHz.

## 5.2. Hyperparameter settings

Following the baseline method, the STFT is computed using a 64 ms sine window with 75% overlap, resulting in a frame period of 16 ms and $F = 513$ unique frequency bins.

The re-implementation of the non-sequential VAE follows the same hyperparameter setting in [10]. Both the encoder and decoder consist of a 128-dimensional fully-connected feed-forward representation layer before mapping to a latent layer of dimension 16, resulting in 171,297 learnable parameters.

For the STCN we use $L = 4$ layers and a filter size $k = 2$ which results in a receptive field of 240 ms. The deterministic representations have dimensions [64, 32, 16, 8], whereas the latent layers have dimensions [32, 16, 8, 4]. The networks which calculate the parameters of the approximate posterior and the prior distribution in each layer, $f_q^{(l)}$ and $f_p^{(l)}$ respectively, are modeled as three stacked layers of 1D convolutions with kernel size 1, mapping from the input dimension $Z_{l+1} + D_l$ to the latent dimension $Z_l$. The observation model uses a TCN of two layers with dimensionality 256, mapping from input dimension $Z = \sum_l Z_l$ to the output dimension $F$. All variances in the latent layers are clamped between 0.001 and 5 and dropout is set to 0.2. The total number of learnable parameters is 325,497 which is less than factor two compared against the non-sequential VAE.

| Method | SI-SDR (dB) | ESTOI | POLQA |
|---|---|---|---|
| VAE [10] | $3.74 \pm 0.21$ | $0.58 \pm 0.01$ | $1.83 \pm 0.03$ |
| STCN | $\mathbf{4.48 \pm 0.30}$ | $\mathbf{0.66 \pm 0.01}$ | $\mathbf{2.20 \pm 0.04}$ |
| Mixture | $0.54 \pm 0.31$ | $0.65 \pm 0.01$ | $2.26 \pm 0.05$ |

Table 1: *Average results and confidence intervals*

We use the Adam optimizer with standard configuration and a learning rate of $10^{-3}$ [34]. For the VAE, we set the batch size to 128, whereas for the STCN to 16. The training takes about 100 epochs until the loss converges. In the training of the STCN we gradually turn on the KL-term $\mathcal{L}_t^{\text{KL}}(\mathbf{x}_t)$ within the first 50 epochs, in order not to collapse into the prior [35].

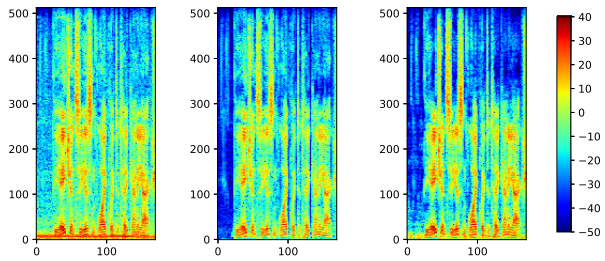
For the optimization of the NMF noise parameters and the time-varying gain, we follow the MCEM algorithm in [10] and set the rank of the NMF to 8. The Metropolis-Hastings algorithm takes the same parameters as in the baseline method.

## 5.3. Results

To measure performance we use the scale-invariant signal-to-distortion ratio (SI-SDR) [36], raw scores of the extended short-time objective intelligibility (ESTOI) with values between 0 and 1 [37], and the perceptual objective listening quality analysis (POLQA) score with values between 1 and 5 [38].

The average results are shown in Table 1. It may be seen that the proposed STCN outperforms the VAE by 0.8 dB in terms of SI-SDR. It is interesting to see that, at the same time, it also outperforms the VAE in terms of ESTOI and POLQA.

Fig. 3 shows spectrograms $\mathbf{S}_{\text{dB}} = 20\log_{10}(|\mathbf{S}|)$ for an example utterance for: a) the mixture signal, b) the corresponding clean speech, and c) the reconstructed clean speech. It may be seen that the proposed method covers speech while attenuating additive noise. Code and audio examples are available online[1].




(a) *Mixture* (b) *Clean* (c) *Reconstructed*

Figure 3: *Magnitude of frequency bins in dB over time frames of an example utterance for: (a) the mixture signal, (b) the corresponding clean speech, and (c) the reconstructed speech.*

# 6. Conclusion

In this work, we proposed a generative speech model based on an STCN for speech enhancement. This approach combines both hierarchical and temporal dependencies of stochastic variables. We evaluated our approach with real recordings of different noise environments. The proposed approach outperforms the non-sequential VAE in terms of signal-to-distortion ratio, and instrumental speech quality and intelligibility.

[1] `https://uhh.de/inf-sp-stcn2020`

# 7. References


[1] E. Vincent, T. Virtanen, and S. Gannot, *Audio Source Separation and Speech Enhancement*. John Wiley & Sons, 2018.

[2] P. C. Loizou, *Speech Enhancement: Theory and Practice*. CRC press, 2013.

[3] R. C. Hendriks, T. Gerkmann, and J. Jensen, *DFT-Domain Based Single-Microphone Noise Reduction for Speech Enhancement – A Survey of the State-of-the-art*. Morgan & Claypool, Jan. 2013.

[4] C. Breithaupt, T. Gerkmann, and R. Martin, "A novel a priori SNR estimation approach based on selective cepstro-temporal smoothing," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2008, pp. 4897–4900.

[5] T. Gerkmann and R. C. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," in *Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4. IEEE, 2011, pp. 1383–1393.

[6] T. Gerkmann, "Bayesian estimation of clean speech spectral coefficients given a priori knowledge of the phase," in *Transactions on Signal Processing*, vol. 62, no. 16. IEEE, 2014, pp. 4199–4208.

[7] E. Vincent, M. G. Jafari, S. A. Abdallah, M. D. Plumbley, and M. E. Davies, "Probabilistic modeling paradigms for audio source separation," in *Machine Audition: Principles, Algorithms and Systems*. IGI global, 2011, pp. 162–185.

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

[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 *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 716–720.

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

[11] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," in *International Conference on Learning Representations ICLR*, Y. Bengio and Y. LeCun, Eds., 2014.

[12] D. J. Rezende, S. Mohamed, and D. Wierstra, "Stochastic backpropagation and approximate inference in deep generative models," in *International Conference on Machine Learning*, 2014, pp. 1278–1286.

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

[14] B. Xia and C. Bao, "Wiener filtering based speech enhancement with weighted denoising auto-encoder and noise classification," *Speech Communication*, vol. 60, pp. 13–29, 2014.

[15] E. Aksan and O. Hilliges, "STCN: Stochastic temporal convolutional networks," in *7th International Conference on Learning Representations (ICLR 2019)*, 2019.

[16] S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," *arXiv preprint arXiv:1803.01271*, 2018.

[17] C. K. Sønderby, T. Raiko, L. Maaløe, S. K. Sønderby, and O. Winther, "Ladder variational autoencoders," in *Advances in neural information processing systems*, 2016, pp. 3738–3746.

[18] C. Févotte, N. Bertin, and J.-L. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis," *Neural computation*, vol. 21, no. 3, pp. 793–830, 2009.

[19] N. Mohammadiha, T. Gerkmann, and A. Leijon, "A new linear MMSE filter for single channel speech enhancement based on nonnegative matrix factorization," in *2011 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2011, pp. 45–48.

[20] A. Liutkus, R. Badeau, and G. Richard, "Gaussian processes for underdetermined source separation," *IEEE Transactions on Signal Processing*, vol. 59, no. 7, pp. 3155–3167, 2011.

[21] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[22] C. Robert and G. Casella, *Monte Carlo Statistical Methods*. Springer Science & Business Media, 2013.

[23] J. Chung, K. Kastner, L. Dinh, K. Goel, A. C. Courville, and Y. Bengio, "A recurrent latent variable model for sequential data," in *Advances in neural information processing systems*, 2015, pp. 2980–2988.

[24] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE transactions on neural networks*, vol. 5, no. 2, pp. 157–166, 1994.

[25] T. Salimans and D. P. Kingma, "Weight normalization: A simple reparameterization to accelerate training of deep neural networks," in *Advances in neural information processing systems*, 2016, pp. 901–909.

[26] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[27] G. Van Den Brink, "Detection of tone pulse of various durations in noise of various bandwidths," *The Journal of the Acoustical Society of America*, vol. 36, no. 6, pp. 1206–1211, 1964.

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[29] A. v. d. 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.

[30] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[31] S. Leglaive, X. Alameda-Pineda, L. Girin, and R. Horaud, "A recurrent variational autoencoder for speech enhancement," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, May 2020.

[32] J. Garofolo, D. Graff, D. Paul, and D. Pallett, "CSR-I (WSJ0) Sennheiser LDC93S6B," *Web Download. Philadelphia: Linguistic Data Consortium*, 1993.

[33] D. B. Dean, A. Kanagasundaram, H. Ghaemmaghami, M. H. Rahman, and S. Sridharan, "The QUT-NOISE-SRE protocol for the evaluation of noisy speaker recognition," in *Interspeech*, 2015, pp. 3456–3460.

[34] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[35] S. R. Bowman, L. Vilnis, O. Vinyals, A. M. Dai, R. Jozefowicz, and S. Bengio, "Generating sentences from a continuous space," *arXiv preprint arXiv:1511.06349*, 2015.

[36] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR – half-baked or well done?" in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 626–630.

[37] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 11, pp. 2009–2022, 2016.

[38] "P.863: Perceptual objective listening quality prediction," International Telecommunication Union, Mar. 2018, iTU-T recommendation. [Online]. Available: https://www.itu.int/rec/T-REC-P.863-201803-I/en