Speech Enhancement Using Bayesian Wavenet

Kaizhi Qian[1]*, Yang Zhang*[1]*, Shiyu Chang*[2]*, Xuesong Yang*[1]*,*
Dinei Florêncio[3]*, Mark Hasegawa-Johnson*[1]

[1]University of Illinois at Urbana-Champaign, USA
[2] IBM Watson Research Center, USA [3]Microsoft Research, USA
{kqian3,yzhan143,xyang45,jhasegaw}@illinois.edu, shiyu.chang@ibm.com, dinei@microsoft.com

Abstract

In recent years, deep learning has achieved great success in speech enhancement. However, there are two major limitations regarding existing works. First, the Bayesian framework is not adopted in many such deep-learning-based algorithms. In particular, the prior distribution for speech in the Bayesian framework has been shown useful by regularizing the output to be in the speech space, and thus improving the performance. Second, the majority of the existing methods operate on the frequency domain of the noisy speech, such as spectrogram and its variations. The clean speech is then reconstructed using the approach of overlap-add, which is limited by its inherent performance upper bound. This paper presents a Bayesian speech enhancement framework, called BaWN (Bayesian WaveNet), which directly operates on raw audio samples. It adopts the recently announced WaveNet, which is shown to be effective in modeling conditional distributions of speech samples while generating natural speech. Experiments show that BaWN is able to recover clean and natural speech.

Index Terms: speech enhancement, Bayesian framework, WaveNet, convolutional neural network, model-based

1. Introduction

Deep learning has been widely used in speech enhancement tasks, because its strong representation power is capable of characterizing complex noise distributions. For example, some works directly predict output spectrum using deep neural networks (DNN) or denoising auto-encoders [1–4]. A series of works [5,6], applied different deep learning architectures to predict ideal ratio masks. Besides, several works performed speech separation using various deep learning architectures [7, 8].

However, these approaches have two major limitations. First, these deep learning algorithms rarely incorporate an explicit prior model for clean speech or a Bayesian framework, which has been shown effective for speech enhancement [9]. While the variability of noise is hardly tractable, the clean speech signal is highly structured, and thus a prior speech model can regularize enhanced speech to become speech-like. Without the speech model, many deep learning algorithms are not generalizable to noises without highly similar characteristics.

On the other hand, existing Bayesian speech enhancement algorithms mostly model speech using simple probability distribution in order to have closed-form solutions. For example, a large body of such works assume HMM-GMM models [10–13] or Laplacian models [14–17]. Others make looser assumptions on kurtosis or neg-entropy of speech distribution [18, 19]. Building a more accurate model for speech becomes a bottleneck for these algorithms, which can potentially be lifted by deep learning.

The second limitation regarding the existing deep learning based approach is that most deep learning algorithms operate on amplitude spectrum, such as short-time Fourier transform or cochleargram. The noisy phase spectrum is directly applied to the enhanced speech without restoring the clean phase spectrum, which may suffer from phase distortion. Also, in some spectral restoration methods, the time domain signal is recovered by overlap-add, which is prone to artifacts and discontinuities. However, applying deep learning directly to speech waveform is difficult, because the high sampling rate requires large temporal memory and receptive field size.

Fortunately, the recently announced WaveNet [20] has demonstrated a strong capability in modeling raw audio waveforms. Its receptive field size is significantly boosted by stacking dilated convolution layers with exponentially increasing dilation rates. Experiments have shown that it is able to generate random babbles with high naturalness. Moreover, WaveNet is probabilistic, which naturally fits into the Bayesian framework.

Motivated by these observations, we propose a Bayesian speech enhancement algorithm using deep learning structures inspired by WaveNet, called the Bayesian WaveNet (BaWN). BaWN directly predicts the clean speech audio samples by estimating the prior distribution and the likelihood function of clean speech using WaveNet-like architectures, which are the two major components of the Bayesian network. It promotes a happy marriage between the Bayesian framework and the deep learning techniques: the former broadens the generalizability for the latter, and the latter improves the model accuracy for the former.

The remainder of the paper is organized as follows. Section 2 describes the architecture of BaWN; section 3 introduces its training scheme; section 4 presents experiments that test its performance; and section 5 concludes the paper.

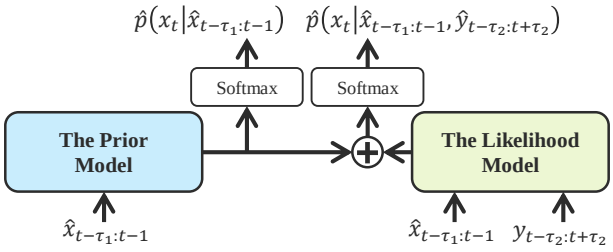
2. The Model Architecture

The problem is formulated within the Bayesian framework. Denote $X_{0:T-1}$ as the random process of the clean speech, which is quantized into Q levels, $q_{0:Q-1}$, via the μ-law encoding [21], so each X_t is a discrete variable. The subscript $0:T-1$ denotes a set with subscripts running from 0 through $T-1$. Denote $Y_{0:T-1}$ as the random process of the observed noisy signal. In this paper, only additive noise is considered, but the framework is generalizable to other types of interferences. Our task is to infer the clean speech $\hat{x}_t$ given a set of noisy observations $Y_{0:T} = y_{0:T}$. For notational ease, probability mass functions will be abbreviated, e.g. $p(X_t = x_t|Y_t = y_t)$ as $p(x_t|y_t)$.

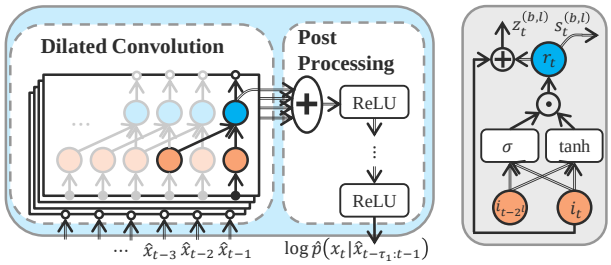
2.1. The Bayesian Framework

We apply a sub-optimal greedy inference scheme for $X_{0:T-1}$. Given inferred values of the past samples $\hat{x}_{0:t-1}$, the inferred

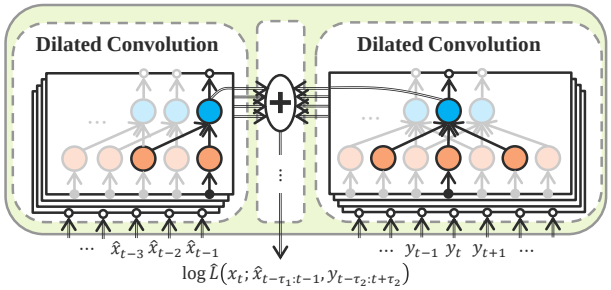
This paper was funded by QNRF grant NPRP 7-766-1-140.



(a) *The general model framework*



(b) *The prior model. The right plot gives a detailed view of a basic convolution unit in the left plot (eq.* (5)*).*



(c) *The likelihood model. The middle module is the post processing module, whose structure is similar to that in (b).*

Figure 1: *The model architecture. Compound arrows denote that the node is multiplied by a weight matrix before sent to the next unit. Circled add and circled dot denote element-wise addition and multiplication respectively. The data path that generates the current output at time t is highlighted.*

value of the current sample, $\hat{x}_t$, is defined as the posterior expectation

$$\hat{x}_t \triangleq \mathbb{E}\left[X_t | X_{t-\tau_1:t-1} = \hat{x}_{t-\tau_1:t-1}, Y_{t-\tau_2:t+\tau_2} = y_{t-\tau_2:t+\tau_2}\right] \tag{1}$$

Here we have made a Markov assumption that the probabilistic dependence of X_t upon variables in the distant past and far future is negligible, when the closer ones, $X_{t-\tau_1:t-1}$ and $Y_{t-\tau_2:t+\tau_2}$, are given. τ_1 and τ_2 denote the range of dependence on $X_{0:T-1}$ and $Y_{0:T-1}$, respectively. Therefore, the following posterior distribution should be evaluated:

$$\begin{aligned} &p(X_t = x_t | X_{t-\tau_1:t-1} = \hat{x}_{t-\tau_1:t-1}, Y_{t-\tau_2:t+\tau_2} = y_{t-\tau_2:t+\tau_2}) \\ \triangleq &p(x_t | \hat{x}_{t-\tau_1:t-1}, y_{t-\tau_2:t+\tau_2}) \\ \propto &p(x_t | \hat{x}_{t-\tau_1:t-1}) \cdot p(y_{t-\tau_2:t+\tau_2} | \hat{x}_{t-\tau_1:t-1}, x_t) \end{aligned} \tag{2}$$

where the $\triangleq$ sign denotes the abbreviation.

Define the likelihood function as

$$L(x_t; \hat{x}_{t-\tau_1:t-1}, y_{t-\tau_2:t+\tau_2}) \triangleq p(y_{t-\tau_2:t+\tau_2} | \hat{x}_{t-\tau_1:t-1}, x_t) \tag{3}$$

Then eq. (2) can be rewritten into

$$\begin{aligned} &p(x_t | \hat{x}_{t-\tau_1:t-1}, y_{t-\tau_2:t+\tau_2}) \\ = &\underbrace{p(x_t | \hat{x}_{t-\tau_1:t-1})}_{\text{prior model}} \cdot \underbrace{L(x_t; \hat{x}_{t-\tau_1:t-1}, y_{t-\tau_2:t+\tau_2})}_{\text{likelihood model}} \end{aligned} \tag{4}$$

The BaWN architecture is based on eq. (4). As shown in Figure 1(a), it consists of two models. The first model is called the prior model, or the speech model, modeling the prior distribution of clean speech signals. For each time t, it takes $\hat{x}_{t-\tau_1:t-1}$ as input, and outputs a Q-dimensional vector of the log estimated pmf $\log \hat{p}(x_t | \hat{x}_{t-\tau_1:t-1})$ up to an unknown constant.

The second model is called the likelihood model, or the noise model, modeling the likelihood function. It takes as inputs $\hat{x}_{t-\tau_1:t-1}$ and $y_{t-\tau_2:t+\tau_2}$, and outputs a Q-dimensional vector of the estimated log likelihood function $\log \hat{L}(x_t; \hat{x}_{t-\tau_1:t-1}, y_{t-\tau_2:t+\tau_2})$ up to an unknown constant.

The two outputs are added and then passed through a softmax nonlinearity. Notice that the exponential function in softmax turns addition into multiplication; the normalization step in softmax removes any unknown constant. Therefore it can be easily shown, from eq. (4), that the output of the softmax nonlinearity is the $p(x_t | \hat{x}_{t-\tau_1:t-1}, y_{t-\tau_2:t+\tau_2})$ of interest. Also, the output of the prior model, passing through a softmax nonlinearity alone, becomes the prior distribution $p(x_t | \hat{x}_{t-\tau_1:t-1})$.

The following two subsections introduce the two models respectively.

2.2. The Prior Model

The prior model replicates the architecture of WaveNet because it performs a similar task. As shown in Figure 1(b), the prior model consists of two modules. The first module is the dilated convolution module, which contains a stack of B_1 blocks with L_1 layers for each. The l-th layer in b-th block is a 1D causal convolution layer through time, with kernel size 2 and dilation rate 2^l. For each time t, it produces two vector outputs—a hidden output $z_t^{(b,l)}$, which is fed into the convolution layer above, and a skip output $s_t^{(b,l)}$, which is directly fed into the second module. The nonlinearity applied is a gated activation unit [22] with residual structure [23]. Formally,

$$f_t^{(b,l)} = \tanh\left(W_{f0}^{(b,l)} i_t^{(b,l)} + W_{f1}^{(b,l)} i_{t-2^l}^{(b,l)} + d_f^{(b,l)}\right) \tag{5a}$$

$$g_t^{(b,l)} = \sigma\left(W_{g0}^{(b,l)} i_t^{(b,l)} + W_{g1}^{(b,l)} i_{t-2^l}^{(b,l)} + d_g^{(b,l)}\right) \tag{5b}$$

$$r_t^{(b,l)} = f_t^{(b,l)} \odot g_t^{(b,l)} \tag{5c}$$

$$z_t^{(b,l)} = i_t^{(b,l)} + W_z^{(b,l)} r_t^{(b,l)} + d_z^{(b,l)} \tag{5d}$$

$$s_t^{(b,l)} = i_t^{(b,l)} + W_s^{(b,l)} r_t^{(b,l)} + d_s^{(b,l)} \tag{5e}$$

where $\sigma(\cdot)$ denotes the sigmoid function; $\odot$ denotes element-wise multiplication; $i_t^{(b,l)}$ denotes the input to this layer,

$$i_t^{(b,l)} = \begin{cases} z_t^{(b,l-1)} & \text{if } l > 0 \\ z_t^{(b-1,L_1-1)} & \text{if } l = 0, b > 0 \\ W_i \hat{x}_t & \text{otherwise} \end{cases} \tag{6}$$

The second module is the post-processing module, which sums all the skip outputs of time t, $s_t^{(0:B_1-1,0:L_1-1)}$, and passes it to a stack of 1×1 convolution (fully connected within time t) layers with ReLU activation. The receptive field size is shown as,

$$\tau_1 = B_1\left(2^{L_1} - 1\right)$$

2.3. The Likelihood Model

The likelihood model is more complex than the prior model. This is because 1) in addition to $\hat{x}_{t-\tau_1:t}$, which is the input to both models, the likelihood model also takes $y_{t-\tau_2:t+\tau_2}$ as input; 2) the prior model is causal, but the likelihood model is non-causal.

To address these complexities, we adapt the original WaveNet structure to that shown in Figure 1(c). The likelihood model also has a dilation convolution module and a post-processing module, but the dilation module now contains two parts. The first part deals with the input $\hat{x}_{t-\tau_1:t}$, and has the same structure as in eqs. (5) and (6). The second part deals with the input $y_{t-\tau_2:t+\tau_2}$, and has almost the same structure, except for two differences. First, the number of blocks and layers within each block is changed to B_2 and L_2 respectively, to accommodate τ_2, which can be different from τ_1. Second, instead of a causal convolution with kernel size 2, this part imposes a non-causal convolution with kernel size 3 to account for future dependency. Formally, eqs. (5a) and (5b) are adapted to

$$f_t^{(b,l)} = \tanh\left(W_{f0}^{(b,l)} i_t^{(b,l)} + W_{f1}^{(b,l)} i_{t-2^l}^{(b,l)} + W_{f-1}^{(b,l)} i_{t+2^l}^{(b,l)} + d_f^{(b,k)}\right) \tag{7a}$$

$$g_t^{(b,l)} = \sigma\left(W_{g0}^{(b,l)} i_t^{(b,l)} + W_{g1}^{(b,l)} i_{t-2^l}^{(b,l)} + W_{g-1}^{(b,l)} i_{t+2^l}^{(b,l)} + d_g^{(b,l)}\right) \tag{7b}$$

The post-processing module in the likelihood model is the same as that in the prior model, except that it sums all the skip outputs from both parts of the dilated convolution module.

3. Training the Model

Since the two models in BaWN have their own specific interpretations, the training scheme should be designed carefully to ensure that the models generate the correct outputs.

3.1. Training the Prior Model

If we replace the input $\hat{x}_{t-\tau_1:t-1}$ with the true clean samples, denoted as $x^*_{t-\tau_1:t-1}$, then the prior model can be trained on clean speech, following a similar paradigm as in WaveNet. Specifically, for each t, given the previous true clean speech, $x^*_{t-\tau_1:t-1}$ as input, the training scheme minimizes the cross entropy between the estimated prior distribution and the empirical distribution. Formally, the training scheme solves the following optimization problem:

$$\max \sum_{t=0}^{T-1} \sum_{i=0}^{Q-1} \mathbb{1}\left\{x_t^* = q_i\right\} \log \hat{p}(X_t = q_i | x_{t-\tau_1:t-1}) \tag{8}$$

where $\mathbb{1}\{\cdot\}$ denotes the indicator function, which equals 1 if the statement in its argument is true and 0 otherwise.

In this paper we only implement the speaker dependent enhancement task. The generalization to speaker independent models will be one of our future directions.

3.2. Training the Likelihood Model

Once the prior model is trained, the likelihood model can be trained by combining both models to estimate the posterior distribution, as indicated by eq. (2). Ideally, we would like to solve

$$\max \sum_{t=0}^{T-1} \sum_{i=0}^{Q-1} \mathbb{1}\left\{x_t^* = q_i\right\} \log \hat{p}(X_t = q_i | \hat{x}_{t-\tau_1:t-1}, y_{t-\tau_2:t+\tau_2}) \tag{9}$$

However, notice that the input of time t contains $\hat{x}_{t-\tau_1:t-1}$, which is a function of the previous time outputs, as shown in eq. (1). Therefore, eq. (9) introduces time recurrence, which causes gradient explosion in practice. An alternative is to replace $\hat{x}_{t-\tau_1:t-1}$ with the true value $x^*_{t-\tau_1:t-1}$ as in prior model training, but this approximation leads to insufficient training, because the model is given too much oracle information about the clean speech.

Our solution is to replace $\hat{x}_{t-\tau_1:t-1}$ with the inferred clean speech produced by the network trained in the *previous iteration*. Denote the previous inferred value as $\hat{x}^{(\text{old})}_{t-\tau_1:t-1}$, then the problem in eq. (9) is reformulated as

$$\max \sum_{t=0}^{T-1} \sum_{i=0}^{Q-1} \mathbb{1}\left\{x_t^* = q_i\right\} \log \hat{p}(X_t = q_i | \hat{x}^{(\text{old})}_{t-\tau_1:t-1}, y_{t-\tau_2:t+\tau_2}) \tag{10}$$

Obtaining the previous inferred value $\hat{x}^{(\text{old})}_{t-\tau_1:t-1}$ can be implemented efficiently using the method in [24].

It should be emphasized that while optimizing for eq. (10), the weights of the prior model should be held fixed to prevent deviation from modeling the prior distribution.

4. Experiments

This section presents experiments that test the performance of the proposed BaWN model. In particular, we will investigate how the prior model improves the generalizability of BaWN to deal with completely unseen and different noises. The ideal ratio mask (DNN-IRM) based model [5] was also implemented as a baseline. Source code can be found at http://tiny.cc/7t5dly.

4.1. Configurations

The three dilated convolutional networks of the WaveNet enhancement model all have 4 blocks of 10 layers, which makes a receptive field size of approximately two to three phones. For each layer, the hidden output has 32 channels and the skip output has 1024 channels. The post-processing modules in both the prior and the likelihood models contain two fully-connected layers, each with 1024 hidden nodes. The clean speech is quantized into 256 levels, so the output dimension is 256.

The training dataset consists of a clean training set (for the prior model) and a noisy training set. The clean training set contains a total of 9700 utterances (19 hours) from audio books played by a female speaker [25]. The noisy training set was created by mixing the 9700 clean utterances randomly with 100 environment noises from [4, 26, 27], including train, airport, restaurant and ring tones. The SNR of the noisy training set is set to two levels: 0dB and -5dB.

There are two test sets, respectively containing 20 and 100 clean utterances of the same speaker randomly selected from another audio book. For the first test set, called the unseen noise test set, 100 noises were selected from a completely different noise dataset [28] in order to test the generalizablity of BaWN, where the types of noises and recording configurations completely differ from that of the training noise dataset. For investigation purpose, the second test set, called the seen noise test set, contains 20 noises drawn from the training noise dataset.

The prior model was trained on all 9700 (19 hours) clean utterances. Due to significantly increased model complexity and the EM-like training procedures, the likelihood model was trained only on 500 (1 hour) utterances from the noisy training set. Though the small sized training data may lead to an insufficiently trained likelihood model, it actually provides a good opportunity to verify the power of the prior model and test the generalizablity of BaWN. For fair comparison, the DNN-IRM

Table 1: *Average SNR, SAR, SDR, STOI of the enhanced utterance using DNN-IRM and BaWN. The first three metrics are measured in decibels (dB), and the STOI is measured in percentage (%). Case indicates the input SNR of the training and testing dataset. Noise indicates whether the noise type is covered by the training set. BaWN stands for Bayesian WaveNet. DIRM stands for DNN-IRM.*

Case	Noise	Model	SNR	SAR	SDR	STOI
0dB	seen	BaWN	22.2	8.53	8.83	85.7
		DIRM	15.6	10.3	12.3	86.4
	unseen	BaWN	22.1	8.37	8.75	84.3
		DIRM	11.9	8.58	12.7	84.8
-5dB	seen	BaWN	21.6	7.15	7.37	81.7
		DIRM	12.2	6.45	8.53	79.0
	unseen	BaWN	20.3	6.65	6.92	80.7
		DIRM	9.20	5.25	8.24	76.6

baseline was trained on the complete noisy training set.

4.2. Objective Evaluation

The performance was measured by the average of SNR, signal-to-artifacts ratio (SAR), signal-to-distortion ratio (SDR), and short-time objective intelligibility (STOI) of the predicted clean utterances. The first three metrics were computed using the BSS-EVAL toolbox [29].

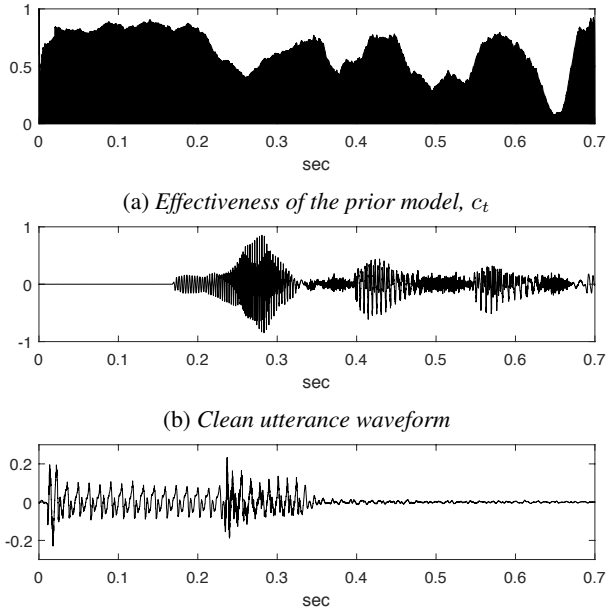
As seen in table 1, the BaWN model outperforms the DNN-IRM model in terms of much higher SNRs. The performance advantage is more significant under the -5 dB case, where BaWN takes the lead in SAR and STOI as well. Also, our model generalizes better to the completely different unseen noises, as the performance drop is smaller. This is remarkable considering that the likelihood model was trained on only one hour of noisy speech and the parameters of the model were not tuned. The prior model has enough knowledge about the distribution of clean speech samples and tends to make non-speech distributions less likely under unseen noises and low SNRs, which helps to make better predictions even if the likelihood model is weak. BaWN achieves slightly lower SDR and, in the 0dB case, SAR, because the sequential inference would occasionally generate impulse noise. Yet this does not weaken our argument for BaWN, considering the inherent negative correlation between the SNR and SAR/SDR, and the huge performance gain in SNR.

4.3. Entropy Analysis

The effectiveness of the prior model under the Bayesian framework can be further visualized and analyzed by computing the entropies of the estimated prior and posterior distribution of each sample. Specifically

$$\begin{aligned} H_t^{(\text{pr})} &= -\sum_{i=0}^{Q} \hat{p}(X_t = q_i | \hat{x}_{t-\tau_1:t-1}) \\ &\quad \cdot \log_2 \hat{p}(X_t = q_i | \hat{x}_{t-\tau_1:t-1}) \\ H_t^{(\text{post})} &= -\sum_{i=0}^{Q} \hat{p}(X_t = q_i | \hat{x}_{t-\tau_1:t-1}, y_{t-\tau_2:t+\tau_2}) \\ &\quad \cdot \log_2 \hat{p}(X_t = q_i | \hat{x}_{t-\tau_1:t-1}, y_{t-\tau_2:t+\tau_2}) \end{aligned} \tag{11}$$

Since the prediction of a sample is more uncertain if the entropy of the corresponding distribution is high, we can conclude



(a) *Effectiveness of the prior model,* c_t

(b) *Clean utterance waveform*

(c) *Noise waveform*

Figure 2: *The prior effectiveness function (eq.* (12)*) of an speech segment, smoothed by a 20-ms moving average filter, with its corresponding utterance and noise.*

that the prior model plays a more important role than the likelihood model at time t if $H_t^{(\text{pr})} < H_t^{(\text{post})}$. Hence we define a prior effectiveness function

$$e_t = \mathbb{1}\left(H_t^{(\text{pr})} < H_t^{(\text{post})}\right) \tag{12}$$

to depict the real-time effectiveness of the prior model. e_t is further smoothed by a 20-ms moving average filter.

Figure 2 shows the smoothed e_t of a test speech segment (a), as well as its corresponding clean speech (b) and noise (c) waveforms. There are two important observations. First, the prior model is more effective when the SNR is low, as can be seen from the segment before 0.25s. This is because when the SNR is high enough, the likelihood model can simply pass noisy observation through, which does not rely much on the prior model.

Second, the prior model is more effective after the onset of vowels or voiced consonants. Accordingly, the likelihood model is more effective during unvoiced consonants or at the onset of speech activities, as can be seen from dips in the effectiveness function at around 0.4s, 0.5s and 0.65s. This is because the voiced speech is well-structured, so the prior model knows what comes next once it recognizes the phone. On the other hand, the prior model is less certain about the unvoiced phones because they are stochastic and can be easily confused with noises.

5. Conclusion

We proposed a WaveNet enhancement model that directly operates on speech waveforms and exploited its generalizability to completely unseen noises. The results showed that our proposed model is able to produce clean speech and outperformed the DNN-IRM model under small-sized training data in terms of generalizability owing to the effectiveness of the prior model.

6. References

[1] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "An experimental study on speech enhancement based on deep neural networks." *IEEE Signal Processing Letters*, vol. 21, no. 1, pp. 65–68, 2014.

[2] D. Liu, P. Smaragdis, and M. Kim, "Experiments on deep learning for speech denoising." in *INTERSPEECH*, 2014, pp. 2685–2689.

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

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

[5] J. Chen, Y. Wang, S. E. Yoho, D. Wang, and E. W. Healy, "Large-scale training to increase speech intelligibility for hearing-impaired listeners in novel noises," *The Journal of the Acoustical Society of America*, vol. 139, no. 5, pp. 2604–2612, 2016.

[6] J. Chen and D. Wang, "Long short-term memory for speaker generalization in supervised speech separation," in *INTERSPEECH*, 2016, pp. 3314–3318.

[7] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Deep learning for monaural speech separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 1562–1566.

[8] F. Weninger, J. R. Hershey, J. Le Roux, and B. Schuller, "Discriminatively trained recurrent neural networks for single-channel speech separation," in *IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, 2014, pp. 577–581.

[9] Z. Ou and Y. Zhang, "Probabilistic acoustic tube: a probabilistic generative model of speech for speech analysis/synthesis." in *AISTATS*, 2012, pp. 841–849.

[10] Y. Ephraim, "A Bayesian estimation approach for speech enhancement using hidden Markov models," *IEEE Transactions on Signal Processing*, vol. 40, no. 4, pp. 725–735, 1992.

[11] H. Sameti, H. Sheikhzadeh, L. Deng, and R. L. Brennan, "HMM-based strategies for enhancement of speech signals embedded in nonstationary noise," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 5, pp. 445–455, 1998.

[12] D. Y. Zhao and W. B. Kleijn, "HMM-based gain modeling for enhancement of speech in noise," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 3, pp. 882–892, 2007.

[13] A. Kundu, S. Chatterjee, A. S. Murthy, and T. Sreenivas, "GMM based Bayesian approach to speech enhancement in signal/transform domain," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2008, pp. 4893–4896.

[14] R. Martin and C. Breithaupt, "Speech enhancement in the DFT domain using Laplacian speech priors," in *International Workshop on Acoustic Echo and Noise Control (IWAENC)*, vol. 3, 2003, pp. 87–90.

[15] T. Lotter and P. Vary, "Speech enhancement by MAP spectral amplitude estimation using a super-Gaussian speech model," *EURASIP Journal on Applied Signal Processing*, vol. 2005, pp. 1110–1126, 2005.

[16] T. Kim, H. T. Attias, S.-Y. Lee, and T.-W. Lee, "Blind source separation exploiting higher-order frequency dependencies," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 70–79, 2007.

[17] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari, "Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 9, pp. 1626–1641, 2016.

[18] B. W. Gillespie, H. S. Malvar, and D. A. Florêncio, "Speech dereverberation via maximum-kurtosis subband adaptive filtering," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 6, 2001, pp. 3701–3704.

[19] K. Kumatani, J. McDonough, B. Rauch, D. Klakow, P. N. Garner, and W. Li, "Beamforming with a maximum negentropy criterion," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 5, pp. 994–1008, 2009.

[20] 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," *arXiv preprint arXiv:1609.03499*, 2016.

[21] "Pulse code modulation (PCM) of voice frequencies," *International Telecommunication Union (ITU)*, 1988.

[22] A. van den Oord, N. Kalchbrenner, L. Espeholt, O. Vinyals, A. Graves, and K. Kavukcuoglu, "Conditional image generation with pixelCNN decoders," in *Advances in Neural Information Processing Systems*, 2016, pp. 4790–4798.

[23] 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.

[24] T. L. Paine, P. Khorrami, S. Chang, Y. Zhang, P. Ramachandran, M. A. Hasegawa-Johnson, and T. S. Huang, "Fast wavenet generation algorithm," *arXiv preprint arXiv:1611.09482*, 2016.

[25] S. King and V. Karaiskos, "The Blizzard Challenge 2013," *Proc. Blizzard Workshop*, 2013.

[26] "Freesound," https://freesound.org/, 2015.

[27] G. Hu, "100 nonspeech sounds," http://web.cse.ohio-state.edu/pnl/corpus/HuNonspeech/HuCorpus.html, 2015.

[28] "FreeSFX," http://www.freesfx.co.uk/, 2017.

[29] C. Févotte, R. Gribonval, and E. Vincent, "BSS_EVAL toolbox user guide–revision 2.0," 2005.