Interspeech 2018
2-6 September 2018, Hyderabad

Speech intelligibility enhancement based on a non-causal Wavenet-like model

Muhammed Shifas PV, Vassilis Tsiaras, Yannis Stylianou

Speech Signal Processing Laboratory (SSPL), University Of Crete, Greece
{shifaspv,tsiaras,yannis}@csd.uoc.gr

Abstract

Low speech intelligibility in noisy listening conditions makes more difficult our communication with others. Various strategies have been suggested to modify a speech signal before it is presented in a noisy listening environment with the goal to increase its intelligibility. A state-of-the art approach, referred to as Spectral Shaping and Dynamic Range Compression (SSDRC), relies on modifying spectral and temporal structure of the clean speech and has been shown to considerably improve the intelligibility of speech in noisy listening conditions. In this paper, we present a non-causal Wavenet-like model for mapping clean speech samples to samples generated by SSDRC. A successful non-linear mapping function has the potential to be used a) in improving the intelligibility of noisy speech and b) in the Wavenet-based speech synthesizers as a model based intelligibility improvement layer. Objective and subjective results show that the Wavenet-based mapping function is able to reproduce the intelligibility gains of SSDRC, while by far it improves the quality of the modified signal compared to the quality obtained by SSDRC.

Index terms: speech intelligibility, Wavenet-like model, sample generation

1. Introduction

In day to day conversations the talker tries to retain the intelligibility factors across the conversation to make speech understandable to the listener. Maintaining the speech intelligibility above a threshold level by adapting to the surrounding noise conditions will turn the listening task easier during the conversations. The articulatory modifications produced by human during the adaptation to the surrounding noise conditions have quite well been studied [1]. A system that is able to simulate this behavior of the human articulatory effort has a great importance on designing effective speech reproduction and in general speech processing systems. For example, in the case of text-to-speech synthesizer operating in competing noise and/or talker scenario such a system will improve the intelligibility of the synthesized speech in order to overcome the disturbance masking. This is an active research area which is widely known as the intelligibility enhancement of speech in noise, or listening enhancement. Improvement of the speech intelligibility in noise can be achieved by several techniques such as boosting the consonants energies [2], or through spectral tilt flattening [3], or formant sharpening and dynamic compression [4]. A state-of-the-art method, referred to as Spectral Shaping and Dynamic Range Compression (SSDRC), has been shown to provide high intelligibility gain in various noisy listening conditions and outperforms other approaches. SSDRC suggests a redistribution of signal energy by applying time-frequency modifications.

Though SSDRC gives excellent improvement in intelligibility when it it is applied on clean speech, the performance drops dramatically when the input signal is noisy. This is because the noise in the noisy speech will also be enhanced, in the sense it will be present. This is because there is not a mechanism in SSDRC to distinguish between noise and speech. It has initially been designed to work on clean speech only. This might be due to modifications applied in the magnitude spectrum during spectral shaping and therefore phase information is ignored. In this work, our target is to design a new method for speech intelligibility enhancement which will have the potential to address the issue to improve the intelligibility of noisy speech. Such a method will be quite applicable in many situations like face to face communications, telecommunications, human-machine interfaces etc, where noisy speech is rather more common than clean speech.

Recently, Wavenet was suggested as a way to generate speech/audio samples through a non-linear autoregressive approach based on deep learning [5]. Also a Wavenet-based approach has been suggested for speech denoising [6]. Inspired by the work on [6], we suggest a Wavenet like approach to map plain speech to SSDRC generated signal using a non-causal Wavenet-like architecture. In short we are looking to define a deterministic function that will be able to map samples of plain speech to those (time-domain) samples generated by SSDRC. Our motivation for such a sample-based non-linear mapping can be also applied on noisy speech. Then we expect at some higher layers a representation of a cleaner version of the input noisy speech will be available to the subsequent higher layers, which will target as their output to be the same as SSDRC-based signals. These target signals have been computed by simply applying SSDRC to the clean version of the input to the network noisy speech. This might also lead to a better quality of modified speech while still intelligibility is maintained. More specifically we will work with a non-causal Wavenet-like architecture exploring, therefore, the conditional dependencies of the sample generated at current time step to the future and past samples of the model input. This modeling of sample dependencies are being implemented through dilated convolution structures. We will refer to this new model as Wavenet-based SSDRC, or shortly wSSDRC. Furthermore, this might help us to easier integrate SSDRC within the latest development in speech synthesis where Wavenet-based approaches are now dominated.

In the sections following, Section 2 briefly explains the SSDRC method. The details of the proposed wSSDRC model for speech intelligibility enhancement are included in Section 3. Experiments with two types of noise conditions will be covered in Section 4, while observations and discussions are provided in Section 5. Finally, Section 6 concludes the paper.

2. Spectral Shaping and Dynamic Range Compression (SSDRC)

SSDRC improves speech intelligibility under various noisy listening conditions by applying spectral shaping and dynamic range compression [7]. It combines properties of Lombard

10.21437/Interspeech.2018-2119

speech, implicit linguistic information and audio processing strategies.

2.1. Spectral shaping (SS)

Spectral shaping is the first stage of the intelligibility enhancement process. It has in total three filters, two of them performs an adaptive to the probability of voicing spectral sharpening that modifies the magnitude of the plain speech. This is followed by a fixed spectral shaping filter, to boost the high frequency components. The module takes the plain speech $x(t)$ as input, in a frame-based processing (frames are of fixed duration) and performs Discrete Fourier Transform (DFT) on each frame to obtain the magnitude spectral components, $X(\omega, t)$. On the adaptive spectral shaping, the local maxima (kind of formants) are sharpened by a spectral sharpening filter $H_s(\omega, t)$ and the high frequency components are being boosted by a pre-emphasis filter $H_p(\omega, t)$ followed. Both of the filters updates their coefficients adaptively on the probability of voicing of individual frames [4]. Hence, the adaptive spectral shaped signal can be written as

$$Y_{aSS}(\omega, t) = H_s(\omega, t)\ H_p(\omega, t)\ X(\omega, t) \tag{1}$$

For boosting the high frequency energies a non-adaptive pre-emphasis filter $H_r(\omega, t)$ is employed to modify the spectra by enhancing the frequency components falling in 1000Hz to 4000Hz by a factor of 12dB, while reduces the frequencies below 500 Hz by 6dB/octave. The spectral shaped signal can be expressed as

$$Y_{SS}(\omega, t) = H_r(\omega, t)\ Y_{aSS}(\omega, t) \tag{2}$$

Inverse Fourier transform and overlap add provides the spectral enhanced speech.

2.2. Dynamic range compression (DRC)

In the dynamic range compression, the idea is to reduce the envelope variation of the speech. This task is achieved through modifying the speech samples in each segment adaptive to the temporal envelopes. DRC is a two step process. In the first stage, envelope is dynamically compressed with recursive smoothing process. The smoothed envelope that is projected to the input output envelope characteristic (IOEC) curve gets the final gain term for the DRC.

Finally, the spectral shaped output from the first module (SS) multiplied by the estimated envelope gains during dynamic range compression (DRC) will provide the final intelligibility enhanced speech by SSDRC.

3. The Proposed wSSDRC Model

Wavenet [5] is a powerful generative approach for the probabilistic modeling of raw audio, which is based on the assumption that speech/audio is a Markov process where the conditional probability for a sample, x_t, given the r previous samples is given by:

$$P(x_t | x_{t-1}, \ldots, x_{t-r}) \tag{3}$$

The Wavenet generates the samples in a way to maximized these conditional probability terms. This conditional mapping has been implemented as an autoregressive network with a stack of residual blocks, Fig.1, where each block contains expert and gate followed the one-dimensional dilated causal convolution. The output of the expert and the gate are being combined via element-wise multiplication. Block, i, computes hidden state vector $h^{(i)}$, Eq.(4), which then being added (due to the residual connections between layers) to the input after a one dimensional convolution, to generate its output $z^{(i)}$.

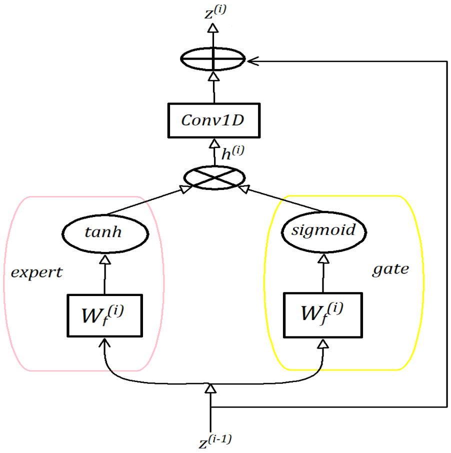

Figure 1: *Residual block to build the model*

$$h^{(i)} = \tanh(W_f^{(i)} * z^{(i-1)}) \odot \sigma(W_g^{(i)} * z^{(i-1)}) \tag{4}$$

$$z^{(i)} = Conv1D(h^{(i)}) + z^{(i-1)} \tag{5}$$

where symbol $*$ denotes convolution and symbol $\odot$ denotes element-wise multiplication.

In this work, similar to Rethage et.al [6], we consider two major modifications in the architecture of Wavenet. First, we use the network as a deterministic mapping, f, from input speech $x = [x_1, \ldots, x_T]$ to an enhanced signal $\hat{y} = [\hat{y}_r, \ldots, \hat{y}_{T-r}]$. Technically, this is done by removing the final softmax layer and adding a layer which projects the output of the post-processing layers to an one-dimensional signal. Also, the compression of the input signal and its 8-bit quantization which are important pre-processing steps in the original Wavenet [5], are not used in this work. Second, instead of considering only the previous r samples of x (receptive of size r), to predict a sample of y at time t, we also consider the next r samples of x, which in essence increased the receptive field size to $2r - 1$. Therefore, the enhanced sample at time $t \in \{r + 1, \ldots, T - r\}$ is predicted as:

$$\hat{y}_t = f(x_{t-r}, \ldots, x_{t-1}, x_t, x_{t+1}, \ldots, x_{t+r}) \tag{6}$$

Fig.2 shows the dependence of the output sample $\hat{y}_t$ on the input samples. As shown in the figure, the dilated convolution structure being used to calculate the activations of the nodes in each block. Which means that the nodes on the i^{th} level in a block ignores the $2^i - 1$ in between samples on the layer below while calculating the response, which is usually been known as the dilation factor of the Wavenet. The skip connections from each blocks are being summed up and processed through a post-processing unit to get the final enhanced samples y_t. The post processing includes two layers of non-causal convolutions having filter width equal to 3 whose output pass through a corresponding ReLu non-linear function, and a one-dimensional convolution, without non-linearity function, which projects to the output one dimensional signal. This model architecture facilitate the generation of set of samples in a single traverse through the structure. when the whole input sequence is available then all output samples can be computed in parallel.

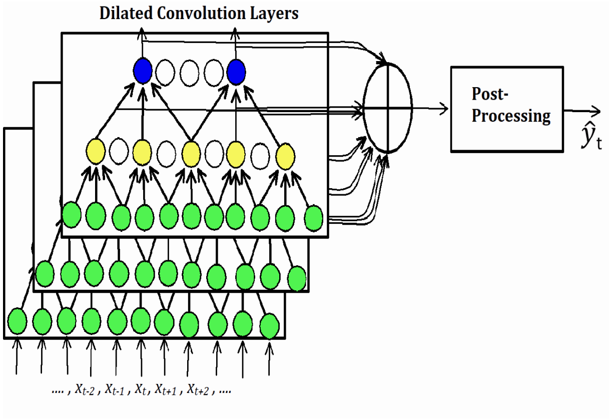
Dilated Convolution Layers
Post-Processing
$\hat{y}_t$
...., x_{t-2}, x_{t-1}, x_t, x_{t+1}, x_{t+2}, ...

Figure 2: *Proposed model architecture*

The model is trained using pairs of time aligned signals $\mathcal{D} = \{(x^{(k)}, y^{(k)}) \mid k \in \{1, \ldots, N\}\}$ by minimizing the average absolute error between a predicted enhanced signal $\hat{y}^{(k)} = f(x^{(k)})$ and the corresponding target enhanced signal $y^{(k)}$.

$$L(x^{(k)}, y^{(k)}) = \frac{1}{T^{(k)} - 2r} \sum_{t=r}^{T^{(k)}-r} |y_t^{(k)} - \hat{y}_t^{(k)}| \qquad (7)$$

where $T^{(k)}$ is the length of signals $x^{(k)}$ and $y^{(k)}$. Therefore, the loss term differs from the actual Wavenet model which had a probability loss function. This is because by removing the final softmax layer from the post-processing stage, we turned the network task to estimate sample error instead of the distribution. The model learns its weights during training by minimizing the above loss. Unlike the actual Wavenet architecture the proposed model doesn't intend to learn the distribution of the output, which makes the conditioning insignificant in the context of this model architecture . Since we have the parallel data samples in hand, the model have specifically designed to generate a set of samples in a shot, rather than individual samples. This gives more momentum for the generation process than the actual Wavenet model and will be practically quite significant for nearly real-time applications.

4. Experimental Setup

For our experiments, clean speech samples from a database provided by the University of Edinburgh has been used [8]. This contains $48kHz$ recorded samples from 28 native English speakers of both genders speaking 400 different sentences. To being fit with our Wavenet-like model the data has been down sampled to $16kHz$. For the noisy conditions, we have considered the speech shaped noise (SSN) and stationary white noise(SWN). For comparison purposes we picked-up the SSDRC system, which is the current best performing system on the enhancement task. For the training of wSSDRC, we used parallel speech data. Which is clean speech samples from the mentioned data set as input to the Wavenet-like network, and the target is being set as the SSDRC modified speech.

Regarding the Wavenet-like network structure, the non-causal convolution filter width is three everywhere in the residual blocks, in which the dilation pattern 1,2,4,..., 512 is repeated three times, resulting in a total of 30 residual blocks and a receptive field length, including the initial non-causal layer and the post-processing layers, of 6145 samples (3072 to the left and 3072 to the right of the current samples). This is sufficient, at $16kHz$ sampling frequency, for modeling the samples dependencies between clean (plain) and SSDRC generated speech. The output from the model is compared with the target through the absolute sample error loss function which has defined on Eq.(7). The loss function are optimized with the Adam optimization algorithm using learning rate 0.0001 and momentum 0.9 [9].

The two systems, SSDRC and the suggested wSSDRC are being compared against intelligibility objective measures and subjective listening test for attesting the quality of the generated speech by the two systems.

For the intelligibility, the Speech Intelligibility Index (SII) has been used. SII captures the intelligibility of a speech signal in noise by looking at the long term average spectral distributions of the energy [10]. For our work we used an extension of the conventional SII by incorporating the temporal characteristics of the noise as well, known as the exSII [11]. All the signals have been normalized, so they have the same loudness before and after modification. The loudness normalization have performed with the recent advanced loudness normalization scheme [12]. The signals have been mixed with SSN and SWN type of noises with Signal to Noise Ratios (SNR) in the range of -5 to 5dB. This tuning of SNR are being done on reference to the plain speech signals.

For measuring the quality performance of the two systems, the SSDRC output and wSSDRC have been compared in a preference test. No added noise have been used. All the signals have been normalized in terms of loudness as it was mentioned above. We have conducted a listening test with 10 subject (non-native English speakers). The subjects have been asked to report back their preferences after listening carefully (using high quality headphones and in a quit office room) samples from both systems: SSDRC and wSSDRC. In total there were nine sets, each having pairs of two utterances one from each system. Same signals have been used in a Mean Opinion Score (MOS) subjective listening test with the same participants. The rating range was 1-bad, 2-poor, 3-fair, 4-good, 5-excellent and participants were presented with some anchor signals for bad, fair and excellent. Anchor signals have been obtained by using the clean plain speech for the excellent rating and two high-pass filtered versions of the original speech for the other two anchor ratings.

5. Observations and Discussion

The results observed from the objective and subjective experiments on both SSDRC and proposed wSSDRC are presented here. An example of speech outputs generated by SSDRC and wSSDRC is provided in Fig. 3. We see that SSDRC and wSSDRC signals are very much similar. The main characteristic is that both produce much lower peak to root mean square (RMS) ratio signals compared to the original plain speech (upper panel of Fig 3). This similar time-domain signature of the processed signals is of course expected as the SSDRC signal is being set as the target of the Wavenet-like model.

The gain in terms of intelligibility score as measured by exSII is shown in the Fig. 4 and Fig. 5 for the two noise types SSN and SWN, respectively. Higher the exSII score better the performance in intelligibility. From the exSII score it is clear that the intelligibility of speech processed through both the SSDRC and the suggested wSSDRC systems have significantly improved compared to that of the unprocessed plain speech. This improvement on intelligibility retained across the SNR

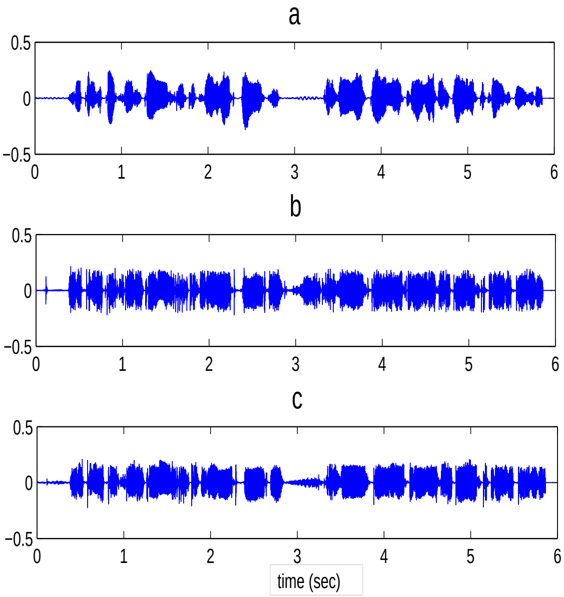

Figure 3: *a)plain speech. b) SSDRC output. c)wSSDRC output*

range. It is worth mentioning that in both types of noise the suggested wSSDRC system can maintain the intelligibility gain at the same level as that of SSDRC. Furthermore, in some cases intelligibility prediction of wSSDRC is even slightly higher to that of SSDRC, when SNR is increasing. For SNR -5dB, we have also informally listened to the noisy signals from SSDRC and wSSDRC and we could confirm that for both systems intelligibility was higher than that of the unprocessed plain speech. We plan to conduct a formal listening test soon.

Table 1: *Preference Test(PT) score in percentage*

Model	PT score
SSDRC	47.3%
wSSDRC	**52.7%**

On the perceptual quality of the speech produced from SSDRC and wSSDRC, we report the results of the preference in Table 1. While wSSDRC was more preferred than SSDRC, the difference is not significant. In the MOS quality test SSDRC have got 3.7 while wSSDRC have got a score of 3.9. The good improvement of quality in case of wSSDRC might be attributed to the sample-by-sample approach Wavenet is using, where at the same time magnitude and frequency information is modified. SSDRC on contrary, only modifies the magnitude information. A further investigation of this cause should be conducted. Also, we will put an effort on improving further the quality of wSSDRC-based speech. One possible avenue to explore is that of AM-FM decomposition of speech. Some speech samples from SSDRC and wSSDRC can be found here [1]

6. Conclusions

In this work, we have suggested a data driven approach for listening enhancement of speech in noise. Specifically, the suggested system has explored the Wavenet approach for modeling the intelligibility patterns on the sample domain of speech. The experimental analysis showed that the suggested wSSDRC system has improved the quality of the processed by a state-of-the-art intelligibility improvement system referred to as SSDRC, while the intelligibility of the speech generated by wSSDRC seems to be at the same levels as the one obtained by SSDRC. The new intelligibility system has the potential to be used as the final layer in state-of-the art Wavenet-based text-to-speech synthesizers for synthesizing intelligible synthetic speech. Furthermore, this way of intelligibility boosting might be applied also in noisy speech. This is our next target.

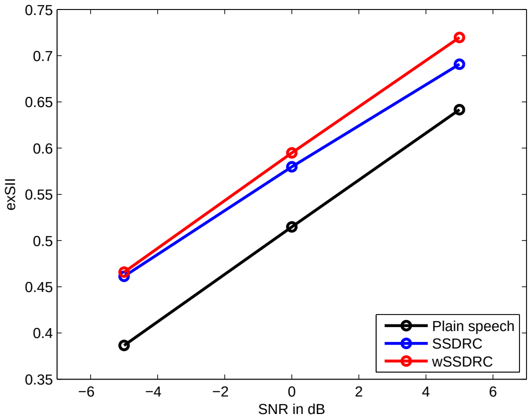

Figure 4: *Intelligibility score as exSII for SSN*

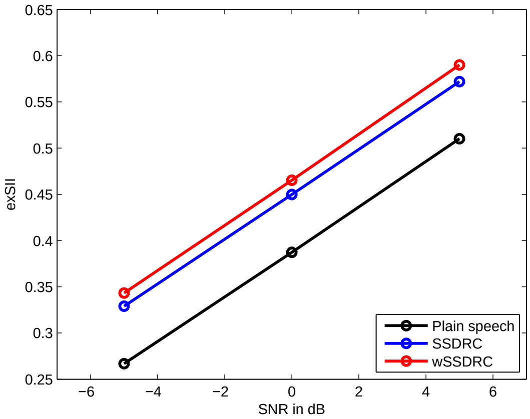

Figure 5: *Intelligibility score as exSII for SWN*

7. Acknowledgements

This work was partly funded by the E.U. Horizon2020 Grant Agreement 675324, Marie Sklodowska-Curie Innovative Training Network, ENRICH.

8. References

[1] B. J. Stanton, L. Jamieson, and G. Allen, “Acoustic-phonetic analysis of loud and lombard speech in simulated cockpit conditions,” in *Acoustics, Speech, and Signal Processing, 1988., International Conference on*. IEEE, 1988, pp. 331–334.

[1] `http://www.csd.uoc.gr/~shifaspv/`

[2] M. D. Skowronski and J. G. Harris, “Applied principles of clear and lombard speech for automated intelligibility enhancement in noisy environments,” *Speech Communication*, vol. 48, no. 5, pp. 549–558, 2006.

[3] Y. Lu and M. Cooke, “The contribution of changes in F0 and spectral tilt to increased intelligibility of speech produced in noise,” *Speech Communication*, vol. 51, no. 12, pp. 1253–1262, 2009.

[4] T.-C. Zorilă and Y. Stylianou, “On spectral and time domain energy reallocation for speech-in-noise intelligibility enhancement,” in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[5] 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.

[6] D. Rethage, J. Pons, and X. Serra, “A wavenet for speech denoising,” *arXiv preprint arXiv:1706.07162*, 2017.

[7] T.-C. Zorila, V. Kandia, and Y. Stylianou, “Speech-in-noise intelligibility improvement based on spectral shaping and dynamic range compression,” in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[8] C. Valentini-Botinhao *et al.*, “Noisy speech database for training speech enhancement algorithms and TTS models,” 2017.

[9] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *CoRR*, vol. abs/1412.6980, 2014. [Online]. Available: http://arxiv.org/abs/1412.6980

[10] P. C. Loizou, *Speech enhancement: theory and practice*. CRC press, 2013.

[11] K. S. Rhebergen and N. J. Versfeld, “A speech intelligibility index-based approach to predict the speech reception threshold for sentences in fluctuating noise for normal-hearing listeners,” *The Journal of the Acoustical Society of America*, vol. 117, no. 4, pp. 2181–2192, 2005.

[12] T.-C. Zorilă, Y. Stylianou, S. Flanagan, and B. C. Moore, “Effectiveness of a loudness model for time-varying sounds in equating the loudness of sentences subjected to different forms of signal processing,” *The Journal of the Acoustical Society of America*, vol. 140, no. 1, pp. 402–408, 2016.