


An Improved Real-Time Noise Suppression Method Based on RNN and Long-Term Speech Information

Baoping Cheng^{1,2}, Guisheng Zhang², Xiaoming Tao¹^a, Sheng Wang², Nan Wu² and Min Chen²

¹Department of Electronic Engineering, Tsinghua University, Beijing, 100084, China

²ChinaMobile (Hangzhou) Information Technology Co., Ltd., Hangzhou, Zhejiang, 311121, China

Keywords: Noise Suppression, Recurrent Neural Network, Long-Term Spectral Divergence, Speech Enhancement.

Abstract: Speech enhancement based on deep learning can provide almost best performance when processing non-stationary noise. Denoising methods that combine classic signal processing with Recurrent Neural Network (RNN) can be implemented in real-time applications due to their low complexity. However, long term speech information is omitted when selecting the features in these methods, which degrades the denoising performance. In this paper, we extend a well-known RNN based denoising method called RNNNoise with the long-term spectral divergence (LTSD) feature. We also limited the amount of noisy speech attenuation to get a better trade-off between noise removal level and speech distortion. Our proposed method outperforms the RNNNoise algorithm by 0.12 MOS points on average in the subjective listening test.

1 INTRODUCTION

Noise suppression (NS) is one of the most active acoustic research area of speech enhancement (SE) since 1970. It aims to improve the performance of human and computer speech interaction by enhancing intelligibility and quality of speech signal which is interfered by ambient noise (Loizou, 2017). Typical NS application scenarios includes but not limited to voice communication, online real-time audio-virtual communications (Valin, 2018), automatic speech recognition, hearing aids, etc.

Conventional signal processing techniques used in NS is to estimate the statistical characteristic of noisy speech signal, and then filter out the noise part or apply a spectral suppression gain in the time-frequency domain (Benesty et al., 2006). Although the conventional NS methods can achieve a good performance in processing stationary noise of speech signal, it cannot cope with non-stationary noise, such as transient noise, whose characteristic is hard to be estimated. In the last decade, deep learning (DL) has been widely used to present the data characteristic. Some researchers tried to use DL to estimate the characteristic of non-stationary noise in NS task (DL Wang et al.,

2018).

In order to deal with both stationary and non-stationary noise, a recent trend of NS research is the development of a hybrid system that combines DL techniques and conventional methods (Xia et al., 2018).

A hybrid systems proposed in (Tu et al., 2018) gives the estimated clean speech and the suppression rule calculated by a long short-term memory (LSTM)-based direct mapping regression model. The suppression rule is a geometric mean of the estimation of the current frame using the conventional NS method and the suppression rule of the previous frame. The first step is used to decrease stationary noise interference. The second step is to efficiently suppress non-stationary noise components. The algorithm proposed in (Coto et al., 2018) follows a similar method with that of Tu (2018). The noisy signal is first processed by the conventional Wiener filter. Then the signal is further processed by a multi-stream approach based on LSTM networks.

In this paper, we investigate the NS performance of the RNNNoise system with additional long-term spectral divergence (LTSD) feature. We also limit the amount of noisy speech attenuation to get a trade-off between noise suppress level and speech distortion.

^a  Xiaoming Tao is the corresponding author of this paper.

The RNNoise system (Valin, 2018) is briefly overviewed in Section 2. The definition of LTSD feature, model architecture and the training and evaluation procedure are present in Section 3. In Section 4, we make the experiments in different scenarios and compare the NS performance of the proposed method with that of RNNoise system. The conclusion is given in Section 5.

2 RNNoise SYSTEM

The work proposed in this paper is based on the RNNoise that combines both conventional and DL techniques, which is implemented by Jean Marc Valin. The architecture of RNNoise system is shown in Figure 1. A Vorbis window with 20 ms duration and 50% overlap is used in the NS process. 42 input features are used in RNNoise to suppress noise. The first 22 features are Bark Frequency Cepstral Coefficients (BFCCs), which are computed by applying the Discrete Cosine Transformation (DCT) on the log spectrum of Opus scale (Valin et al., 2016). Next 12 features are the first and second ordertemporal derivatives of the first 6 BFCCs. The following 6 features are the first 6 coefficients computed by the DCT on the pitch correlation across frequency bands. The last 2 features are the pitch period and a spectral non-stationary metric respectively.

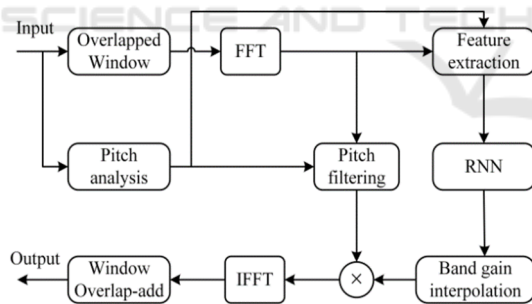


Figure 1: Architecture of RNNoise system

The input of RNNoise is 48 kHz full-band audio signal. The consecutive audio signal is divided into many frames by overlapped window, and then transformed into frequency domain by FFT. Using the outputs of pitch analysis and FFT, 22 features can be obtained, which are the BFCCs. RNN is used to calculate an ideal ratio mask (IRM) for 22 triangular bands derived from the Opus scale. The 22 band gains can be applied to the DFT magnitudes of each window after an interpolation.

Before the IFFT operation, a comb filter defined at the pitch period, also called a pitch filter, is applied

to each window. The aim of the pitch filter is to remove noise components among pitch harmonics in voice segments.

The processed signal is the multiplication of the outputs of pitch filter and band gain interpolation. Then processed signal is converted by IFFT to obtain the time domain signal. Finally, an overlap-add (OLA) method is used to produce the denoised signal. In practical application, the OLA method consecutively process each frame when it arrives.

3 PROPOSED METHOD

In order to deal with the non-stationary noise, the LTSD feature is attached to the 42 features used in RNNoise, since the LTSD feature is useful for discriminating speech and non-speech signal. We use a long-term speech spectrum window instead of instantaneous one to track the spectral envelope and compute LTSD by estimating of the long-term spectral envelope (LTSE).

3.1 Definitions of the LTSD Feature

Let $x(n)$ be a noisy speech signal which is divided into overlapped frame segments. Its amplitude spectrum for the frame l at k band is defined as $X(k, l)$. The N -order LTSE is defined as

$$LTSE_N(k, l) = \max\{X(k, l + j)\}_{j=-N}^{j=+N} \quad (1)$$

The average noise spectrum magnitude for the frame l at k band is defined as $S(k, l)$. The deviation of the N -order LTSE respect to $S(k, l)$ for the band, $k = 0, 1, \dots, NFFT-1$, is defined as N -order LTSD, is described as

$$LTSD_N(l) = 10 \log_{10} \left(\frac{1}{NFFT} \sum_{k=0}^{NFFT-1} \frac{LTSE_N^2(k, l)}{S^2(k, l)} \right) \quad (2)$$

The average noise spectrum magnitude value is known in the training phase and its calculation formula in the evaluation is given by

$$S(k, l) = \begin{cases} \alpha S(k, l-1) + (1-\alpha) \hat{S}(k, l), & \text{if non-speech segment is detected} \\ S(k, l-1), & \text{otherwise} \end{cases} \quad (3)$$

$$\hat{S}(k, l) = X(k, l) \cdot \hat{m}(k, l) \quad (4)$$

$\hat{m}(k,l)$ is the k frequency bin denoising gain computed by the RNN at frame l . The speech and non-speech segment detection is also performed by the RNN which is introduced in Section 3.2. To get a good trade-off between noise reduction and computation complexity, N is set to 6 (Ram et al., 2003).

3.2 Architecture and Implementation

We first extract the 42 features presented in (Valin, 2018) and extend them with the additional LTSD feature to evaluate our proposed system. We pass this data along with the input audio files to the evaluation tool which computes, interpolates and applies the opus scale gains along with a pitch filter to the audio file.

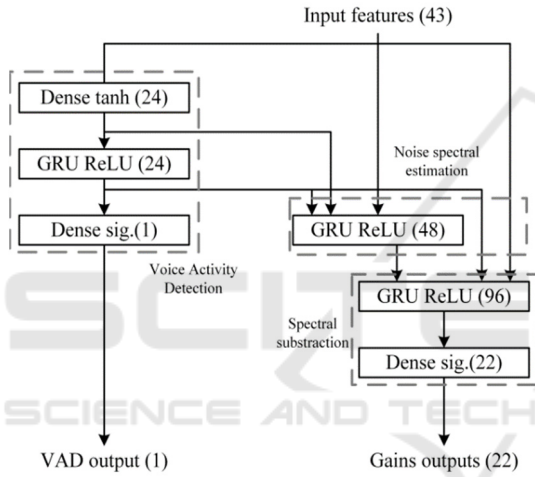


Figure 2: Deep recurrent neural network topology

The architecture of the neural network used follows the original RNNNoise architecture with the difference that the input layer created a tensor whose size is modified to fit that of the increased number of features. The detailed topology is shown in Figure 2 and the network contains 215 units and 4 hidden layers.

As described in (Valin, 2018), the entire network architecture is designed so that it follows the structure of most conventional noise suppression methods. The denoising system can be divided into three modules: a voice activity (VAD), a noise spectral estimation module and a spectral subtraction module. Each module includes a recurrent layer and applying GRU.

VAD module contributes significantly to the training phase by assisting the system differentiate noise

from speech. It also outputs a voice activity flag used to compute the average noise spectrum in the evaluation phase.

The full training and evaluation phase is visualized in Figure 3. We first use Sound eXchange (SoX) to concatenate and convert the input clean speech and noise to RAW format. Then we apply the appropriate tool to produce the samples used for training, by mixing clean speech and noise samples in different SNRs, and extract the original 42 features. After, we process the training samples to extract the additional LTSD feature using formula (1) and (2).

Training

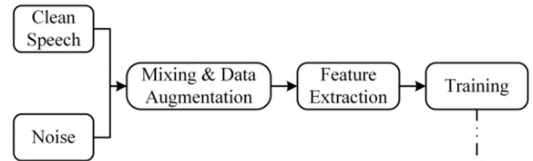


Figure 3: Training and Evaluation overview

We train the proposed system using the Keras toolchain with Tensorflow backend. Both reference and proposed network are trained through the course of 160 epochs with 8 steps. We set the learning rate to 0.001 by applying the Adam optimizer. We use the loss function (3) (as proposed by Vallin), where m is the ground truth NS gain, \hat{m} is the mask calculated by the RNN, $\gamma = 1/2$ is a parameter that tunes the NS aggressiveness and K is the number of bands, set to 22 in our training. During training, both systems are fed with 6000000 audio frames, each with a non-overlapping 10 ms duration.

$$L(m, \hat{m}) =$$

$$\frac{1}{N} \left(10 \cdot \sum_{i=1}^K (\min(m_i + 1, 1) \cdot (10 \cdot (m_i - \hat{m}_i)^4 + (\sqrt{\hat{m}_i} - \sqrt{m_i})^\gamma - 0.01 \cdot m_i \cdot \log(\hat{m}_i))) \right) - \frac{1}{2} \cdot \sum_{i=1}^K (2 \cdot |m_i - 0.5| \cdot m_i \cdot \log(\hat{m}_i)) \quad (5)$$

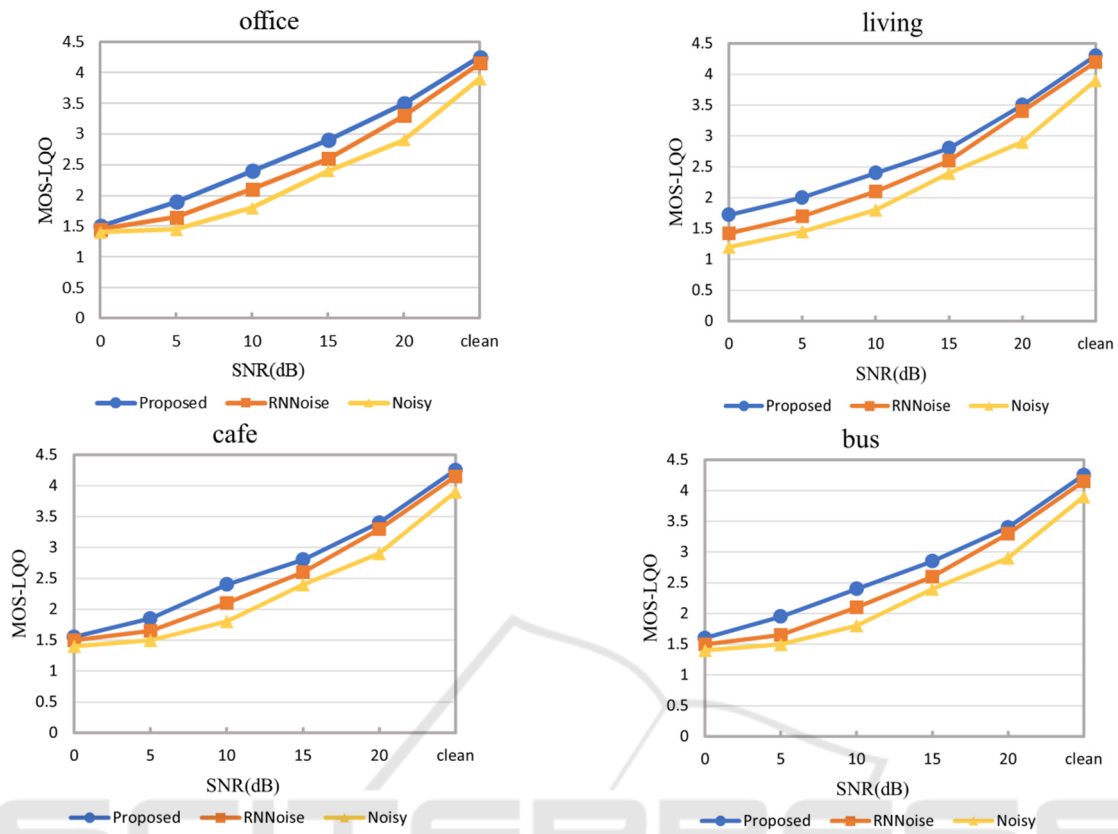


Figure 4: PESQ MOS-LQO quality evaluation for living, office, cafe, and bus noise environment

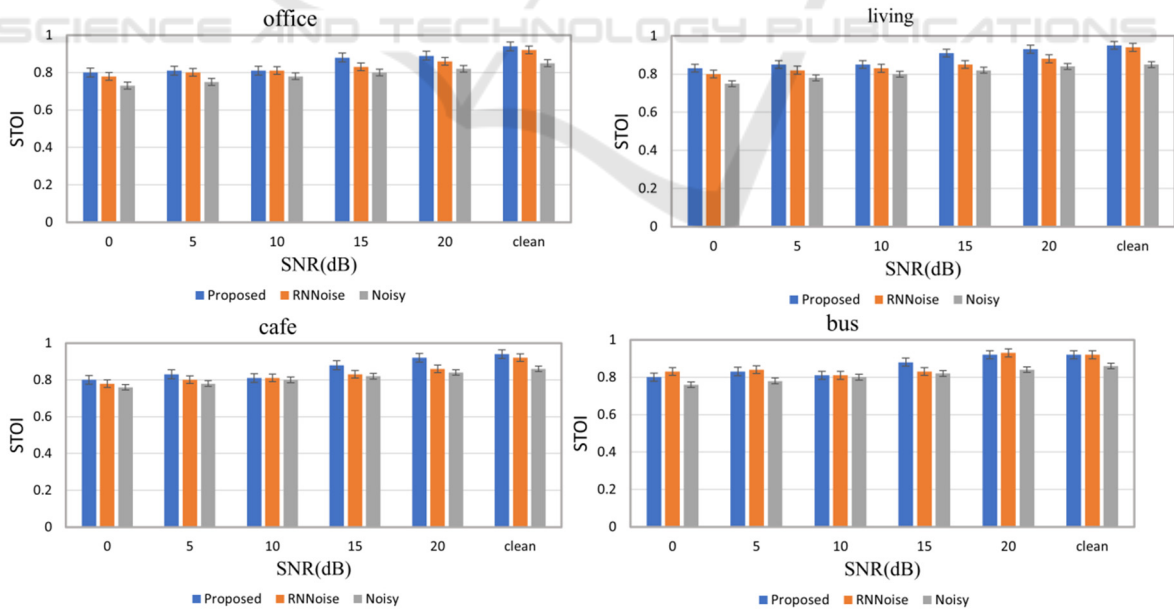


Figure 5: STOI in different acoustical environments under various SNR levels

We apply the clean speech datasets included in the Edinburgh Datasets (Botinhao, 2017) which is comprised of audio recordings, including 28 English speakers (14 men and 14 women), sampled at 48 kHz, to train our proposed system. For noise audios, we used a subset of the noise environments available in the DEMAND datasets (Thiemann et al., 2013). These noise environments were then not included in the test set. The DEMAND datasets include noise recordings corresponding to six distinct acoustic scenes (Domestic, Nature, Office, Public, Street and Transportation), which are further subdivided in multiple more specific noise sources (Thiemann et al., 2013). Note that while we used clean speech and noise included in the Edinburgh Datasets, the samples used for training the systems are not the noisy samples found in the noisy speech subset of the Edinburgh Datasets, but rather samples mixed using the method described in (Valin, 2018).

The model generalization is achieved by data augmentation. Since cepstrum mean normalization is not applied, the speech and noise signal are filtered independently for each training example through a second order filter as

$$H(z) = \frac{1 + r_1 z^{-1} + r_2 z^{-2}}{1 + r_3 z^{-1} + r_4 z^{-2}} \quad (6)$$

where each of $r_1 \dots r_4$ are following the uniform distribution in the $[-3/8, 3/8]$ range. We vary the final mixed signal level to achieve robustness to the input noisy speech signal. The amount of noisy speech attenuation is also limited to get a better trade-off between noise removal level and speech distortion.

The test set used is the one provided in the Edinburgh Datasets, which has been specifically created for SE applications and consists of wide-band (48 kHz) clean and noisy speech audio tracks. The noisy speech in the set contains four different SNR levels (2.5dB, 7.5dB, 12.5dB, 17.5dB). The clean speech tracks included in the set are recordings of two English language speakers, a male and a female. As for the noise recordings that were used in the mixing of the noisy speech tracks, those were selected from the DEMAND database. More specifically, the noise profiles found in the testing set are:

- Office: noise from an office with keyboard typing and mouse clicking
- Living: noise inside a living room
- Cafe: noise from a cafe at a public square
- Bus: noise from a public bus

The selected evaluation metrics is of great importance in the effort of regular evaluation of one system. In order to evaluate our system, we used a metric

that focuses on the intelligibility of the voice signal (STOI) ranges from 0 to 1 and a metric that focuses on the sound quality (PESQ) ranges from -0.5 to 4.5, with higher values corresponding to better quality.

4 EXPERIMENTAL RESULTS AND DISCUSSIONS

It was appropriate to present our results in a comparison between the reference RNNNoise system and the proposed method that makes use of the LTSD feature. As seen in Figure 4, it becomes apparent that the proposed method has better performance in most acoustic scenarios and in all SNR levels, especially in lower SNRs, comparing the two methods with the PESQ quality metric. Our proposed method outperforms the RNNNoise algorithm by 0.12 MOS points on average. Similarly, examining the STOI intelligibility measure, as depicted in Figure 5, it is indicated that the proposed method also has better intelligibility performance.

We observe a noticeable improvement in performance than RNNNoise method. Having also compared several spectrograms of both methods, it was observed that in general the proposed method does indeed subtract more noise components.

Having taken these results into consideration, it is demonstrated that more detailed research is required in future work to reduce speech distortion and promote noise removal level for different application scenarios.

Firstly, we find that adding more hidden layers indeed be beneficial for our proposed system. Given that we provide the system with more and diverse input information, the RNN might be able to better use the proposed features with additional hidden layers.

Secondly, studying samples processed by our extended system, we speculate that the system could benefit from changing how aggressively the noise suppression occurs. This can be achieved by fine-tuning the value of the γ parameter in the loss function (3), keeping in mind that smaller γ values lead to more aggressive suppression. According to our experiment, setting $\gamma = 1/2$ is an optimal balance.

Finally, we believe that further research can be done regarding the performance of our proposed systems as the training datasets increases in size and diversity.

5 CONCLUSIONS

In this paper we extend and improve a hybrid noise suppression method. We proposed LTSD feature which we believed would be beneficial to the denoising process and regard it as one input to the network. We describe our implementation for training the system with extended input features and make a comparison against the reference RNN trained by the same training parameters and datasets. We make a discussion about our findings from this process, concluding that the extra LTSD feature have obvious positive effect on NS performance. In the future, we will explore the further improvement of the base system by tuning hidden layers, loss function and datasets.

REFERENCES

- Loizou, P. C. (2007). *Speech enhancement: theory and practice*, pages 795-809. CRC press.
- Benesty, J., Makino, S., & Chen, J. (Eds.), (2006). *Speech enhancement*, pages 345-360. Springer Science & Business Media.
- Shifas, M. P., Adiga, N., Tsiaras, V., & Stylianou, Y. (2020). A non-causal FFTNet architecture for speech enhancement. *arXiv preprint arXiv:2006.04469*.
- Lim, J. S., & Oppenheim, A. V. (1979). Enhancement and bandwidth compression of noisy speech. *Proceedings of the IEEE*, 67(12):1586-1604.
- Wang, D., & Chen, J. (2018). Supervised speech separation based on deep learning: An overview. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(10): 1702-1726.
- Xia, Y., & Stern, R. M. (2018). A Prior SNR Estimation Based on a Recurrent Neural Network for Robust Speech Enhancement. In *INTERSPEECH*, pages 3274-3278.
- Valin, J. M. (2018, August). A hybrid DSP/deep learning approach to real-time full-band speech enhancement. In *2018 IEEE 20th international workshop on multimedia signal processing (MMSP)*, pages 1-5. IEEE.
- Ram Rez, J. Segura, J. C., Ben Tez, C., ángel de la Torre, & Rubio, A. (2004). Efficient voice activity detection algorithms using long-term speech information. *Speech Communication*, 42(3-4), pages 271-287.
- Valin, J. M., Maxwell, G., Terriberry, T. B., & Vos, K. (2016). High-quality, low-delay music coding in the opus codec. *arXiv preprint arXiv:1602.04845*.
- Tu, Y. H., Tashev, I., Zarar, S., & Lee, C. H. (2018, April). A hybrid approach to combining conventional and deep learning techniques for single-channel speech enhancement and recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2531-2535. IEEE.
- Coto-Jimenez, M., Goddard-Close, J., Di Persia, L., & Ruffner, H. L. (2018, July). Hybrid speech enhancement with wiener filters and deep lstm de-noising autoencoders. In *2018 IEEE International Work Conference on Bioinspired Intelligence (IWOB1)*:2531-2535. IEEE.
- Valentini-Botinhao, C. (2017). Noisy speech database for training speech enhancement algorithms and tts models.
- Thiemann, J., Ito, N., & Vincent, E. (2013, June). The diverse environments multi-channel acoustic noise database (demand): A database of multichannel environmental noise recordings. In *Proceedings of Meetings on Acoustics ICA2013*, 19(1):035081. Acoustical Society of America.