TRUTH-TO-ESTIMATE RATIO MASK: A POST-PROCESSING METHOD FOR SPEECH ENHANCEMENT DIRECT AT LOW SIGNAL-TO-NOISE RATIOS

Bohan Chen[1]*, He Wang*[1]*, Yue Wei*[2]*, and Richard H.Y. So*[1,3]

[1] Hong Kong University of Science and Technology Shenzhen Research Institute, Shenzhen, China
[2] Incus Company Limited, Hong Kong SAR
[3] Department of Industrial Engineering and Decision Analytics,
Hong Kong University of Science and Technology, Hong Kong SAR
{bhchen, hwanga, rhyso}@ust.hk, yweiaj@connect.ust.hk

ABSTRACT

This study proposes a bi-directional recurrent neural network (Bi-RNN) post-processing method for speech enhancement (SE) at low signal-to noise ratios (SNR). Current speech enhancement solutions performed badly under low SNR situations. Loizou and Kim proposed a solution to reduce speech distortion errors in time-frequency (T-F) domain but it requires the knowledge of ground truth. As ground truth is unknown in real-life applications, the current study proposes to use a Bi-RNN to implement Loizou and Kim's solution as a post-processing method for SE engines. Our solutions do not require prior knowledge of ground truth. The effectiveness of the proposed method is investigated with a spectral subtraction (SS) SE engine, a non-negative matrix factorization (NMF) SE engine, and a deep neural network ideal ratio mask (DNN-IRM) SE engine, under matched/mis-matched noise and different SNR conditions. Experimental results demonstrate that the proposed post-processing method effectively improved both perceptual evaluation of speech quality (PESQ) and short-time objective intelligibility (STOI) for all of these SE engines, especially at low SNR conditions.

Index Terms— Post-processing, Speech enhancement, Low signal-to-noise ratio, Bi-directional recurrent neural network

1. INTRODUCTION

Single channel speech enhancement aims to reduce background noise and improve the quality and intelligibility of degraded speech recorded from single microphone. It has been studied actively over the past decades because of its wide range of applications. Conventional unsupervised algorithms such as spectral subtraction (SS) [1, 2], minimum mean-square estimator (MMSE) [3, 4] are still active in both laboratory and industry because of their flexibility and simplicity. On the other hand, for the supervised methods, efforts are made to make use of prior information from a large amount of speech and noise data. These algorithms, such as hidden Markov model (HMM) [5] and non-negative matrix factorization (NMF) [6], have shown better performance against non-stationary noise. In the past few years, deep learning based supervised methods have become the mainstream [7]. Benefited from the rapid rise in deep learning, it has continuously updated the state-of-the-art performance of speech enhancement.

In general, the aforementioned algorithms work well at high SNR (i.e., SNR $\geq$ 6 dB in this study) environment, especially under known stationary background noise. When the SNR is low (i.e., SNR $<$ 6 dB in this study), however, speech enhancement becomes difficult in that even the state-of-the-art algorithms fail to produce ideal enhanced signal [8]. The reasons behind this decrease of performance are complicated, possibly including the ineffectiveness of voice activity detection [9], the deviation of phase [10], the difficulty to access to the true noise spectrum [8], etc. It is not good news because human listeners also have no difficulty in understanding words at high SNR level [11]. In other words, current speech enhancement solutions have bad performance when the human listeners start to get in trouble with noise.

It has been shown that some typical SE problems, such as musical noise and phase reconstruction, can be remedied by introducing additional post-processing algorithms[12, 13]. In this paper, we propose a Bi-RNN model based post-processing method to detect and modify bins, which underestimate noise, from the enhanced signal in T-F domain. Experimental results demonstrate that the proposed method improves the tested SE engines, especially at low SNR conditions.

2. TRUTH-TO-ESTIMATE RATIO MASK POST-PROCESSING

In [8], Loizou and Kim claimed that under-estimated noises (i.e., T-F units) introduce severe speech distortion. Therefore, it is better to throw these chunks rather than keep them, es-

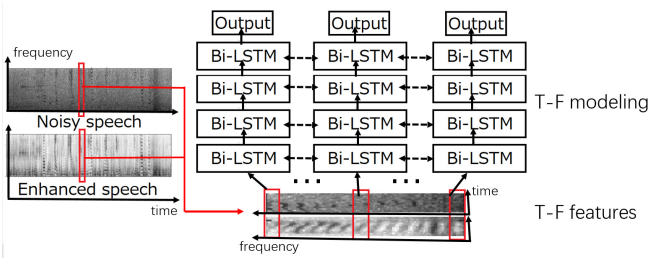

Fig. 1. Diagram of the proposed Bi-RNN model. Magnitude spectra of the enhanced speech together with the corresponding noisy magnitude spectra are set to be the input features. A 4-hidden-layer Bi-RNN is used to learn the $\mathrm{TERM}(t,f)$. Each hidden layer contains a bi-directional long short term memory (Bi-LSTM) layer.

pecially at low SNR conditions where under-estimation frequently occurs. To prove their idea, they proposed a constraint to modify enhanced signal. Given the ground truth of the magnitude spectrum of the clean speech $X(t,f)$, at time t and frequency f, the magnitude spectrum of modified enhanced speech $\hat{X}_M(t,f)$ can be written as:

$$\hat{X}_M(t,f) = \mathrm{TERM}(t,f) \times \hat{X}(t,f), \tag{1}$$

where $\hat{X}(t,f)$ denotes the enhanced magnitude spectrum. $\mathrm{TERM}(t,f)$ denotes truth-to-estimate ratio mask, which can be written as:

$$\mathrm{TERM}(t,f) = \begin{cases} 1, & \text{if } \frac{X(t,f)}{\hat{X}(t,f)} \geq LC \\ 0, & \text{else} \end{cases} \tag{2}$$

By setting the local criterion LC to 1 and $1/2$, Loizou and Kim showed that "large gains in intelligibility were obtained by human listeners, even by spectral-subtractive algorithms".

We notice that TERM is similar to masks that used to train deep neural network based SE engines. In fact, TERM can be interpreted as a special ideal binary mask (IBM) [14], which treats enhanced magnitude spectrum (i.e., $\hat{X}(t,f)$) as noisy magnitude spectrum (i.e., $Y(t,f)$) and sets the local criterion of "SNR" to $+\infty$ (when $LC = 1$). We hence consider TERM can be realised using neural network, without knowing the ground truth. In this paper, we proposed to use Bi-RNN to model $\mathrm{TERM}(t,f)$. The diagram of the proposed model is shown in Fig. 1. Magnitude spectra of the enhanced speech $\hat{X}(t-m,f),\ldots,\hat{X}(t,f),\ldots,\hat{X}(t+n,f)$, together with the corresponding noisy magnitude spectra $Y(t-m,f),\ldots,Y(t-m,f),\ldots,Y(t+n,f)$ are set to be the input features. A 4-hidden-layer Bi-RNN is used to learn the aforementioned $\mathrm{TERM}(t,f)$. Each hidden layer contains a bi-directional long short term memory (Bi-LSTM) layer with 256 hidden units and hyperbolic tangent (i.e., tanh) activation function, and a dropout layer with dropout probability set to be 0.2. Note that as Bi-RNN is introduced to model not the correlation between frames but frequency bins, it has a fixed number of "frequency" steps which is equal to half of the number of short-time Fourier transform points (i.e., 128 in our experiment).

In practice, LC is set to be 1 rather than $1/2$, because we found it has better performance in our model. Note that $LC = 1$ makes TERM a very strict constraint which tends to reject every T-F unit that under-estimates noise. We use 2 forward frames, 2 backward frames (i.e., $n = m = 2$) and the current frame as our input. We use posterior probabilities (i.e., sigmoidal output of the Bi-RNN with binary training targets) as a soft mask for T-F unit modification, because it is found to have better modification performance compared to hard binary mask (i.e., Eq. (2)). Similar suggestion has also been reported in [15].

3. EXPERIMENTS AND RESULTS

3.1. SE engines

3.1.1. spectral subtraction (SS)

SS is one of the most basic speech enhancement algorithms which has been developed for decades. The performance of SS depends both on its noise power spectral density (PSD) estimator and subtraction rule. In this paper, we use the SS engine in VOICEBOX [16] which use the noise PSD estimator and subtraction rule proposed by Martin [2, 17].

3.1.2. non-negative matrix factorization (NMF)

NMF is a widely used matrix factorization algorithm based on the observation that lots of the real world data only contains non-negative data vectors. In this paper, we build our NMF SE engine similar to [18] which uses the speech and noise power spectrum estimated by NMF as the parameters of a Wiener filter.

3.1.3. deep neural network ideal ratio mask (DNN-IRM)

Instead of a hard label on each T-F unit (e.g., IBM), in [15], the authors proposed to use the ideal ratio mask (IRM) as a soft version of the IBM. In IRM based SE, the enhanced magnitude spectrum $\hat{X}(t,f)$ can be written as: $\hat{X}(t,f) = \mathrm{IRM} \times Y(t,f)$, where

$$\mathrm{IRM}(t,f) = \left(\frac{X(t,f)^2}{X(t,f)^2 + N(t,f)^2}\right)^{\beta},$$

where $N(t,f)$ denotes the magnitude spectrum of noise, and the tunable parameter β is commonly set to be 0.5 [15]. In this study, we use 3 hidden layers deep neural network (DNN) to learn the IRM. Each hidden layer contains 2048 hidden units and scaled exponential linear units (i.e., selu) activation function and a dropout layer with dropout probability set to be 0.2.

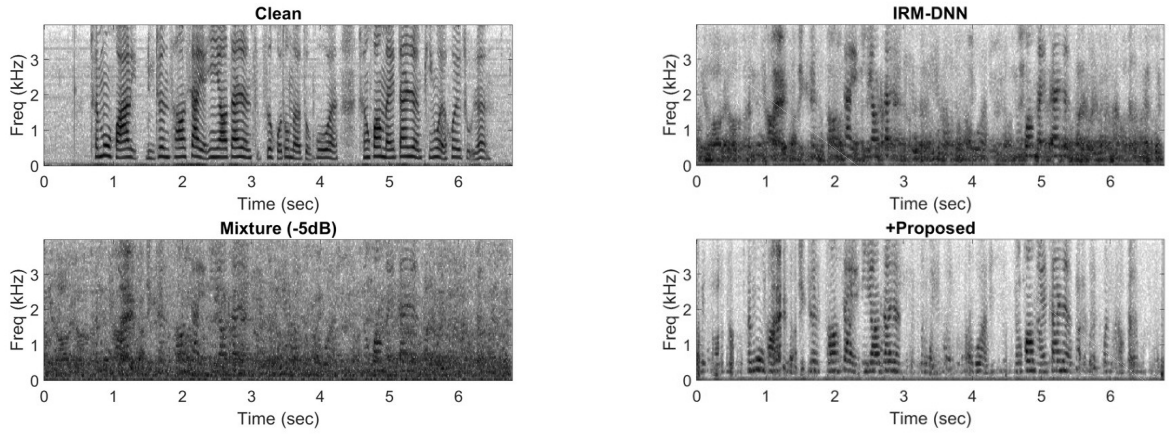

Fig. 2. Examples of experimental results. (a) Spectrograms of clean signal and Mixture (babble, -5 dB SNR) signal. (b) Spectrograms of DNN-IRM enhanced signal and post-processed DNN-IRM enhanced signal.

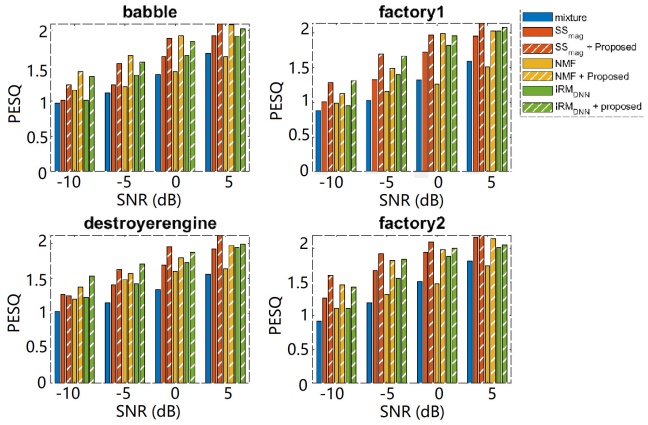

Fig. 3. PESQ measure values under four noise conditions. Blue bars denote PESQ values of mixture (unprocessed) signals. Filled bars and diagonal striped bars denote PESQ values before and after proposed post-processing, respectively. Red bars, yellow bars and green bars denote PESQ values of SS SE, NMF SE and DNN-IRM SE, respectively.

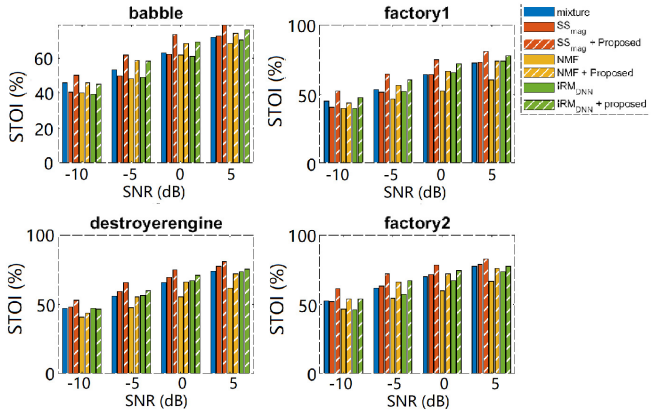

Fig. 4. STOI measure values under four noise conditions. Blue bars denote STOI values of mixture (unprocessed) signals. Filled bars and diagonal striped bars denote STOI values before and after proposed post-processing, respectively. Red bars, yellow bars and green bars denote STOI values of SS SE, NMF SE and DNN-IRM SE, respectively.

3.2. Experimental setup

We use 1000 randomly chosen utterances from the THCHS30 [19] training set as training utterances for the supervised SE engines (i.e., DNN-IRM and NMF), another 500 randomly chosen utterances from the same training set as training utterances for the proposed Bi-RNN model, 100 randomly chosen utterances from the THCHS30 test set as our test utterances. We use 2 noises (i.e., "babble" and "factory1") from the NOISEX dataset [20] as training (both supervised SE engines and the proposed method) and test (matched) noises. 2 additional noises (i.e., "destroyerengine" and "factory2") are used for test-only (mis-matched) noise. To create the training sets, we use random cuts from the first half of each noise to mix with the training utterances around -5 and 0 dB SNR[1]. The test mixtures are generated by mixing random cuts from the last half of each noise with the test utterances around -10, -5, 0 and 5 dB SNR. In our experiment, aforementioned algorithms only estimate the magnitude/power spectrum of the clean speech, the phase of the noisy speech is directly used for reconstruction.

Perceptual Evaluation of Speech Quality (PESQ) score

[1]We add a random value from $(-1, 1)$ to the SNR to increase variability

[21] and Short-Time Objective Intelligibility score (STOI) [22], both of which have been shown to be highly correlated with human subjective speech intelligibility score, are chosen to be the objective measurements of the experiment.

3.3. Results

The spectrograms of the clean speech, its mixture (unprocessed) with babble noise at -5 dB SNR and the enhanced speech from DNN-IRM SE engine before and after proposed post-processing method are shown in Fig. 2. The spectrogram of the post-processed enhanced speech is significantly more similar to the original clean speech with more speech-liked details, which can hardly been found before post-processing. On the other hand, as a result of its aggressive denoising policy, our proposed method also shows a tendency of "white-washing" the spectrogram. In subjective listening, it sounds like residual musical noise which is typically produced by IBM SE engine. It makes sense because the proposed method can be interpreted as a special IBM for enhanced speech.

The PESQ scores and the STOI scores in different experimental conditions are shown in Figs. 3 and 4, respectively. Levels of relative improvement (i.e., in terms of % improvement) are reported below, as we believe this will provide readers a more generic understanding of the improvement. The average improvements under matched noise (i.e., babble and factory1) through different SNR conditions are 16.95% and 18.20% (for PESQ score and STOI score, respectively, the same below), for SS SE engine, 32.94% and 16.91% for NMF SE engine, and 13.45% and 12.21% for DNN-IRM SE engine. The average improvements under mismatched noise (i.e., destroyerengine and factory2) through different SNR conditions are 10.73% and 9.42% for SS SE engine, 21.85% and 16.70 for NMF SE engine, and 11.81% and 7.5% for DNN-IRM SE engine.

In general, the proposed post-processing method shows consistent improvement in objective measurements compared with the original enhanced signal, especially at low SNR conditions. When the SNR increases, the improvement of the proposed method decreases. It is expectable, because TERM is designed to detect under-estimated noises which are more likely to appear at low SNR conditions. As with other supervised algorithms, the proposed method also shows dependency on training data, its performance significantly decreases when dealing with unseen noise, especially at "destroyerengine" condition. As factory2 noise is not entirely unseen with respect to factory1 noise, the real training data dependency of the proposed method is probably stronger than what we show in our experiment.

4. CONCLUSION

In this paper, we propose to use Bi-RNN based TERM as a post-processing mask aiming at removing under-estimated noises for SE engines. The generalizability and effectiveness of the proposed method are investigated with three different SE engines, under matched/mis-matched noise and different SNR conditions. Experimental results demonstrate that the proposed method improved the performance of all SE engines on objective measures (PESQ and STOI), especially at low SNR conditions.

The basic assumption behind the proposed method is that, for SE even the worst over-estimation of noise (i.e., set the T-F unit to 0) is better than under-estimation of noise. Although it has been proven to be effective by [8] and our experiment, it is obvious that the assumption does not always hold. Consequently, the proposed method has an upper bound of improvement caused by its chunk throwing policy. Likewise, the proposed method is also not fit to process enhanced signals which have limited under-estimation chunks (e.g., at a very high SNR condition).

Although TERM is introduced as a detector for under-estimated noises in this paper, it can also serve as a detector for over-estimated noises given some modification. Extending our method to include over-estimated noise detection has potential to further improve the overall performance. This should further increase the general applicability of this post-processing method. Moreover, as TERM can be interpreted as a special form of IBM, other learning targets designed for SE (e.g., IRM, SMM) as well as their corresponding learning algorithms (e.g., CNN, GAN) with various specialities can also be modified into a post-processing engine for SE. The new framework of integrating SE engines showed in this study can be further extended to include these algorithms in the future.

5. ACKNOWLEDGEMENT

The authors would like to thank HKUST Fund of Nanhai (Grant No. FSNH-18FYTRI01); the Science, Technology and Innovation Commission of Shenzhen Municipality for partially supporting the work under project No. JCYJ20170413173515472.

6. REFERENCES

[1] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," in *Transactions on Acoustics, Speech, and Signal Processing*. IEEE, 1979, vol. 27(2), pp. 113–120.

[2] R. Martin, "Spectral subtraction based on minimum statistics," in *Proc. EUSIPCO, Edinburgh*, Sept. 1994, vol. 27(2), pp. 1182–1185.

[3] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," in *Trans Acoustics, Speech, and Signal Processing*. IEEE, 1984, vol. 32(6), pp. 1109–1121.

[4] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," in *Trans Acoustics, Speech, and Signal Processing*. IEEE, 1985, vol. 33(2), pp. 443–445.

[5] H. Sameti, H. Sheikhzadeh, L. Deng, and R. L. Brennan, "Hmm-based strategies for enhancement of speech signals embedded in nonstationary noise," in *Transactions on Speech and Audio processing*. IEEE, 1998, vol. 6(5), pp. 445–455.

[6] N. Mohammadiha, P. Smaragdis, and A. Leijon, "Supervised and unsupervised speech enhancement using nonnegative matrix factorization," in *Transactions on Audio, Speech, and Language Processing*. IEEE, 2013, vol. 21(10), pp. 2140–2151.

[7] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," in *Transactions on Audio, Speech, and Language Processing*. IEEE, 2018, vol. 26(10), pp. 1702–1726.

[8] P. C. Loizou and G. Kim, "Reasons why current speech-enhancement algorithms do not improve speech intelligibility and suggested solutions," in *Transactions on Audio, Speech, and Language Processing*. IEEE, 2010, vol. 19(1), pp. 47–56.

[9] P. K. Ghosh, A. Tsiartas, and S. Narayanan, "Robust voice activity detection using long-term signal variability," in *Transactions on Audio, Speech, and Language Processing*. IEEE, 2010, vol. 19(3), pp. 600–613.

[10] S. R. Chiluveru and M. Tripathy, "Low snr speech enhancement with dnn based phase estimation," in *International Journal of Speech Technology*, 2019, vol. 22(283), pp. 1381–2416.

[11] G. A. Miller, "The masking of speech," in *Psychological bulletin*, 1947, vol. 44(2), pp. 105–129.

[12] N. Madhu, C. Breithaupt, and R. Martin, "Temporal smoothing of spectral masks in the cepstral domain for speech separation," in *International Conference on Acoustics, Speech, and Signal Processing*. IEEE, 2008, pp. 45–48.

[13] Mayer F., Williamson D., Mowlaee P., and Wang D. L., "Impact of phase estimation on single-channel speech separation based on time-frequency masking," in *Journal of the Acoustical Society of America*, 2017, pp. 4668–4679.

[14] D. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech separation by humans and machines*. Springer, Boston, MA, 2005, pp. 181–197.

[15] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," in *Transactions on Audio, Speech, and Language Processing*. IEEE, 2014, vol. 22(12), pp. 1849–1858.

[16] M. Brookes, "Voicebox: Speech processing toolbox for matlabs," .

[17] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," in *Transactions on Speech and Audio Processing*. IEEE, 2001, vol. 9(5), pp. 504–512.

[18] S. M. Kim, J. H. Park, H. K. Kim, S. J. Lee, and Y. K. Lee, "Non-negative matrix factorization based noise reduction for noise robust automatic speech recognition," in *International Conference on Latent Variable Analysis and Signal Separation*, 2012, pp. 338–346.

[19] D. Wang, X. Zhang, and Z. Zhang, "Thchs-30 : A free chinese speech corpus," 2015.

[20] A. Varga and H. J. Steeneken, "Assessment for automatic speech recognition: Ii. noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems," in *Speech Communication*, 1993, vol. 12(3), pp. 247–251.

[21] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *International Conference on Acoustics, Speech, and Signal Processing*. IEEE, 2001, pp. 749–752.

[22] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," in *Transactions on Audio, Speech, and Language Processing*. IEEE, 2011, vol. 19(7), pp. 2125–2136.