# Leveraging the perceptual metric loss to improve the DEMUCS system in speech enhancement

Qi-Wei Hong, Chi-En Dai, Hui-Chun Hsu, Zong-Tai Wu, Jeih-weih Hung

National Chi Nan University
No. 1, University Road
Puli Township, Nantou County, Taiwan
{s108323024, s108323060, s108323058, s110323503}@mail1.ncnu.edu.tw, jwhung@ncnu.edu.tw

**Abstract**

This study aims to improve the source separation technique, DEMUCS, by revising the respective loss function. DEMUCS, developed by Facebook Team, is built on the Wave-U-Net and consists of convolutional layer encoding and decoding blocks with an LSTM layer in between. The applied loss function in DEMUCS contains wave-domain L1 distance and multi-scale short-time-Fourier-transform (STFT) loss.

We present to revise the original loss by considering the perceptual metric scores, including perceptual speech quality (PESQ) and short-time objective intelligibility (STOI). The new loss function becomes a weighted sum of the original loss and the losses of STOI and PESQ, hoping to highlight the perceptual quality of the enhanced utterances.

According to the preliminary experiments conducted on the VoiceBank-DEMUCS task, the DEMUCS network with the modified loss function provides the noise-corrupted utterances with superior objective perceptual metric scores (PESQ and STOI). These results indicate that the presented work benefits DEMUCS in speech enhancement performance.

**Key words:** speech enhancement, DEMUCS, STFT, loss function, PESQ, STOI

## I. Introduction

Speech enhancement (SE) aims to optimize input speech utterances' quality or intelligibility mainly by alleviating the embedded noise and interference. Most conversational speech signals contain some form of noise distortion that hinders understanding, such as crowd noise, keyboard sounds, and dogs barking. Therefore, SE is essential for the applications or tasks that receive speech signals for further employment, such as hearing aids, speech recognition, and speaker recognition. Conventional SE algorithms have used statistical modeling for speech or noise, and they often operate in the short-time domain of speech to reduce the delay of processing. However, these SE methods fail to deal with non-stationary noise widely in real scenarios.

Notably, the rapid development of deep neural networks (DNN) and deep learning algorithms in the recent decade has contributed to a remarkable evolution in SE technologies [1, 2, 3]. These DNN-based SE methods are more successful in dealing with non-stationary noise scenarios than their conventional counterparts. A variety of DNN structures, like fully connected, convolutional, and recurrent neural networks, together with their variants, can serve as the fundamental scheme of an SE algorithm.

Among DNN-based SE methods, the DEMUCS network [4, 5] proposed by the Facebook team is quite effective. It follows an encoder-decoder architecture consisting of convolutional layers and U-net skip connections. Accordingly, this study investigates DEMUCS by changing its loss function that monitors its convergence and attempting to improve its SE performance further. We employ two objective evaluation metrics, perceptual speech quality (PESQ) [6] and short-time objective intelligibility (STOI) [7], adding them to the original loss function to monitor and determine when the network's learning needs to terminate. Preliminary experiments conducted with the VoiceBank-DEMAND task reveal that the presented strategy can further promote the SE performance of DEMUCS.

## II. DEMUCS

The original DEMUCS is designed for multiple source separation. It consists of convolutional encoder and decoder layers with U-net skip connections and a sequence modeling network to revise the encoders' output. When applying DEMUCS to monophonic speech enhancement, we set the number of channels for the input and output of DEMUCS to one.For the details of network arrangements of DEMUCS, one can refer to [4, 5].

The loss function to be minimized in DEMUCS network learning consists of the L1 loss over the waveform and the multiresolution STFT loss over the spectrogram magnitudes. We express the used loss function between the length-$T$ enhanced time waveform $\boldsymbol{y}$ and its clean duplicate $\tilde{\boldsymbol{y}}$ as follows:

$$L_{DEMUCS}(\boldsymbol{y},\tilde{\boldsymbol{y}}) = \frac{1}{T}\|\boldsymbol{y}-\tilde{\boldsymbol{y}}\|_1 + \sum_i L_{stft}^{(i)}(\boldsymbol{y},\tilde{\boldsymbol{y}}) \qquad (1)$$

where $\|.\|_1$ is the L1 norm, the first term on the right-hand side is the L1 loss over the time-domain waveform, and the superscript "$(i)$" is the index of the STFT loss functions with a specified parameter set (hop size and window length). Furthermore, the STFT loss in Eq. (1) contains two parts: the spectral convergence (*sc*) loss and the magnitude (*mag*) loss as follows:

$$L_{stft}(\boldsymbol{y},\tilde{\boldsymbol{y}}) = L_{sc}(\boldsymbol{y},\tilde{\boldsymbol{y}}) + L_{mag}(\boldsymbol{y},\tilde{\boldsymbol{y}}), \qquad (2)$$

where

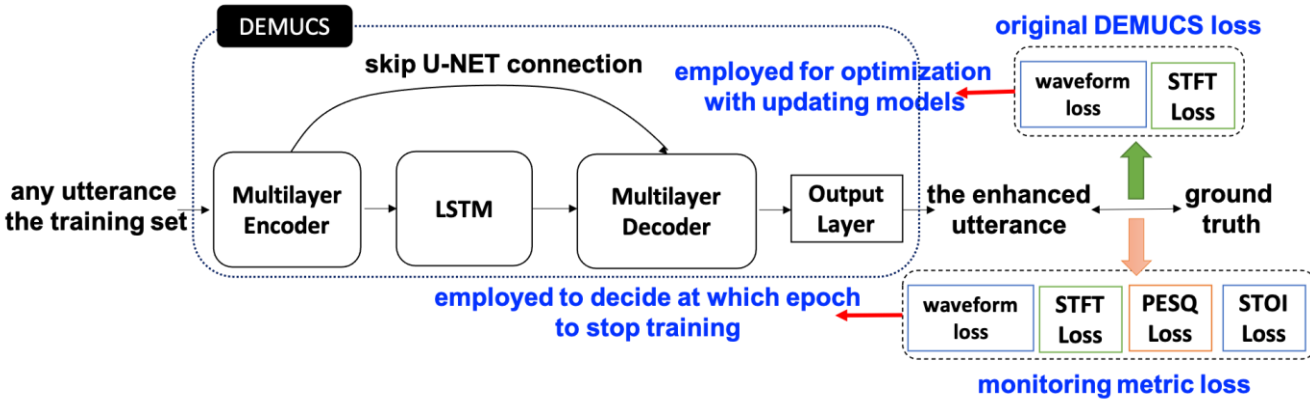


Figure 1. The DEMUCS system with the presented monitoring metric loss

$$L_{sc}(\boldsymbol{y},\tilde{\boldsymbol{y}}) = \frac{\| |STFT(\boldsymbol{y})| - |STFT(\tilde{\boldsymbol{y}})| \|_F}{\|STFT(\boldsymbol{y})\|_F}, \quad (3)$$

and

$$L_{mag}(\boldsymbol{y},\tilde{\boldsymbol{y}}) = \frac{1}{T}\left\| \boldsymbol{log}\left(\frac{|STFT(\boldsymbol{y})|}{|STFT(\tilde{\boldsymbol{y}})|}\right)\right\|_1, \quad (4)$$

where $\|.\|_F$ is the Frobenius norm.

## III. Proposed method

Observing the loss function employed in the DEMUCS network training, we think it comprehensively measures the discrepancy between the enhanced speech and its original noise-free duplicate. Such a loss function, once minimized, can reduce the distortion from the original noise source and the network simultaneously. However, it does not explicitly consider the perceptual quality and intelligibility of enhanced utterances. Therefore, this study presents to further leverage two objective metrics for SE: PESQ and STOI, which measure an utterance's quality and intelligibility. However, since the introduction of PESQ and STOI to the loss function would inevitably increase the computation load of the network learning of DEMUCS, here we propose a "lightweight" revision, which is depicted in Figure 1. The main idea is to use the fusion of the original loss (L1 waveform loss plus multi-resolution STFT loss) and perceptual metric loss as the monitoring metric loss. When the monitoring metric loss for the validation set reaches its minimum at a particular number of epochs in training, we select the DEMUCS network obtained at that epoch number to enhance the test utterances.

Note that the parameter update in DEMUCS still entirely depends on the original loss, while we use the monitoring metric loss to decide when to terminate the update. Performing this way avoids pursuing the gradient of the perceptual metric loss function, which might be complicated or even unavailable. In addition, because the reduction of waveform and spectral discrepancy as in the original loss usually positively correlates with the increase of perceptual metrics, we think using the monitoring metric loss would not seriously deteriorate the ultimate DEMUCS network.

The proposed monitoring metric loss for DEMUCS training is expressed as follows:

$$L(\boldsymbol{y},\tilde{\boldsymbol{y}}) = (1-\alpha-\beta)L_{DEMUCS}(\boldsymbol{y},\tilde{\boldsymbol{y}}) + \alpha f(PESQ(\boldsymbol{y},\tilde{\boldsymbol{y}})) + \beta g(STOI(\boldsymbol{y},\tilde{\boldsymbol{y}})), \quad (5)$$

where $PESQ(\boldsymbol{y},\tilde{\boldsymbol{y}})$ and $STOI(\boldsymbol{y},\tilde{\boldsymbol{y}})$ are the PESQ and STOI metric scores of the enhanced utterance $\boldsymbol{y}$ with respect to its clean duplicate $\tilde{\boldsymbol{y}}$, respectively, $f(.)$ and $g(.)$ are in general monotonically decreasing functions (to reflect the loss characteristics), and $\alpha$ and $\beta$ are two weight factors. Furthermore, unless otherwise specified, we have

$$f\big(PESQ(\boldsymbol{y},\tilde{\boldsymbol{y}})\big) = 0.45 - PESQ(\boldsymbol{y},\tilde{\boldsymbol{y}}) \quad (6)$$

and

$$g(STOI(\boldsymbol{y},\tilde{\boldsymbol{y}})) = 1.0 - STOI(\boldsymbol{y},\tilde{\boldsymbol{y}}), \quad (7)$$

because the upper bounds of PESQ and STOI scores are 0.45 and 1.0, respectively.

## IV. Experimental Setup

We conduct evaluation experiments with the VoiceBank-DEMAND task, in which the utterances and noises are from the VoiceBank [8] and DEMAND [9] corpus, respectively. There are 11,572 utterances for training and 824 utterances for testing. The training set is produced by 28 speakers, added with noise with ten types at four SNRs (0, 5, 10 and 15 dB). The test set is from 2 speakers with five noise types at four SNRs (2.5, 7.5, 12.5 and 17.5 dB). In addition, a validation set is used, which contains 742 utterances.

The causal DEMUCS structure is learned for 300 epochs with the batch size of 32, with some other parameter settings $U = 4, S = 4, K = 8, L = 5$, and $H = 48$ used in [5]. Following Eqs. (1) and (5), we train the DEMUCS network with multiple epochs at different assignments of weight factors. Among the DEMUCS networks corresponding to different epochs, we choose the one that gets the lowest monitoring metric loss to evaluate the test set. We use the PESQ [6] and STOI [7] metrics to evaluate the objective quality and intelligibility of the enhanced utterances in the test set.

## V. Experimental Results and Discussions

First of all, we examine how the functions, $f$ and $g$, of PESQ and STOI in the presented monitoring metric loss influence the DEMUCS performance in SE, and the corresponding results are listed in Table 1. From this table, we have the following observations.

1. All of the settings in Table 1 do not provide any improvement in PESQ and STOI relative to the DEMUCS baseline, probably because the improper settings of the functions $f(.)$ and $g(.)$ and weight parameters $\alpha$ are $\beta$.
2. When we set the two functions $f(.)$ and $g(.)$ are monotonically increasing (such as $f(x) = 10x, f(x) = x, g(x) = 10\,x,$ and $g(x) = 100x$), the SE performance is significantly poor, which confirms our assumption that these metric-related functions have to be negatively correlated with the loss during the training.
3. Setting $f(PESQ) = \mathbf{4.5} - PESQ$ and $g(STOI) = \mathbf{1} - STOI$ achieves better results than setting $f(PESQ) = -PESQ$ and $g(STOI) = -STOI$ . This implies using nonnegative functions would be a better choice to determine the monitoring metric loss.

Table 1. The PESQ and STOI scores of the test set for VoiceBank-DEMAND task achieved by DEMUCS and the presented method with different assignments of functions $f$ and $g$ for PESQ and STOI monitoring metric loss

| | PESQ | STOI |
|---|---|---|
| DEMUCS baseline ($\alpha = \beta = 0$) | 2.9227 | 0.9466 |
| $f(PESQ) = -PESQ$ ($\alpha = 0.33, \beta = 0$) | 2.2193 | 0.9258 |
| $f(PESQ) = \mathbf{4.5} - PESQ$ ($\alpha = 0.33, \beta = 0$) | 2.8500 | 0.9454 |
| $f(PESQ) = \mathbf{10}PESQ$ ($\alpha = 0.33, \beta = 0$) | 1.7352 | 0.8953 |
| $f$(PESQ)=PESQ (α=0.33, β= 0) | 1.7242 | 0.9054 |
| $g(STOI) = -STOI$ ($\alpha = 0, \beta = 0.33$) | 2.7969 | 0.9452 |
| $g(STOI) = 1 - STOI$ ($\alpha = 0, \beta = 0.33$) | 2.9247 | 0.9478 |
| $g(STOI) = 10STOI$ ($\alpha = 0, \beta = 0.33$) | 1.7527 | 0.8701 |
| $g(STOI) = 100STOI$ ($\alpha = 0, \beta = 0.33$) | 1.7438 | 0.8959 |

Next, we fix the two functions, $f$ and $g$, to be as Eqs. (3) and (4) and then tune either of the two weight parameters, $\alpha$ and $\beta$, in Eq. (2) to see the corresponding effect, which experimental results are listed in Tables 2 and 3. We have several observations from these two tables:

1. By properly setting the weight factors, the proposed monitoring loss can provide DEMUCS with significant improvement in SE. For example, setting $\alpha = 0.005$ and $\beta = 0$ increases PESQ and STOI by 0.0243 (from 2.9227 to 2.9471) and 0.0005 (from 0.9466 to 0.9471), and setting $\alpha = 0$ and $\beta = 0.67$ increases PESQ and STOI by 0.0243 (from 2.9227 to 2.9540) and 0.0005 (from 0.9466 to 0.9472).
2. It is noteworthy that adding the PESQ loss alone (with $\alpha \neq 0$ and $\beta = 0$) might improve both PESQ and STOI, and it is also the case for adding STOI loss alone (with $\alpha = 0$ and $\beta \neq 0$). Moreover, the optimal PESQ (2.9540) occurs by adding STOI loss alone with $\beta = 0.67$.
3. As for the case with the addition of PESQ loss ($\alpha \neq 0$), decreasing the value of $\alpha$ can result in better PESQ and STOI, while it is almost the opposite for the case with the addition of STOI loss. A larger value of $\beta$ improves PESQ and STOI more significantly.
4. The possible improvement for PESQ is more significant than that for STOI possibly because the current PESQ score has much room for improvement while the obtained STOI is close to its upper bound (1.0).

Table 2. The PESQ and STOI scores of the test set for VoiceBank-DEMAND task achieved by DEMUCS and the presented method using PESQ monitoring metric loss ($f(PESQ) = 4.5 - PESQ$) with different values of weighting factor $\alpha$

| | | PESQ | STOI |
|---|---|---|---|
| DEMUCS baseline ($\alpha = \beta = 0$) | | 2.9227 | 0.9466 |
| $\beta = 0$ (No STOI is involved) | $\alpha = 0.33$ | 2.8500 | 0.9454 |
| | $\alpha = 0.10$ | 2.9105 | **0.9467** |
| | $\alpha = 0.05$ | 2.8664 | 0.9456 |
| | $\alpha = 0.01$ | **2.9326** | **0.9469** |
| | $\alpha = 0.0075$ | **2.9451** | **0.9471** |
| | $\alpha = 0.005$ | **2.9471** | **0.9471** |

Table 3. The PESQ and STOI scores of the test set for VoiceBank-DEMAND task achieved by DEMUCS and the presented method using STOI monitoring metric loss ($g(STOI) = 1.0 - STOI$) with different values of weighting factor $\beta$

| | | PESQ | STOI |
|---|---|---|---|
| DEMUCS baseline ($\alpha = \beta = 0$) | | 2.9227 | 0.9466 |
| $\alpha = 0$ (No PESQ is involved) | $\beta = 0.25$ | 2.9265 | **0.9469** |
| | $\beta = 0.33$ | **2.9247** | **0.9478** |
| | $\beta = 0.50$ | **2.9294** | **0.9475** |
| | $\beta = 0.57$ | **2.9203** | **0.9471** |
| | $\beta = 0.67$ | **2.9540** | **0.9472** |
| | $\beta = 0.69$ | **2.9500** | 0.9466 |
| | $\beta = 0.82$ | **2.9286** | **0.9473** |

Finally, we evaluate the presented method with the spectrogram of utterances for visual comparison. Fig. 2 depicts the spectrograms of an utterance from the VoiceBank dataset at different conditions. Comparing Figs. 2(a)(b) it is obvious that noise causes significant distortion in the spectrogram. The spectrogram enhanced with the presented method shown in Fig 2(c) reveals that the noise distortion existing in Fig. 2(b) has been greatly removed.

(a) clean utterance

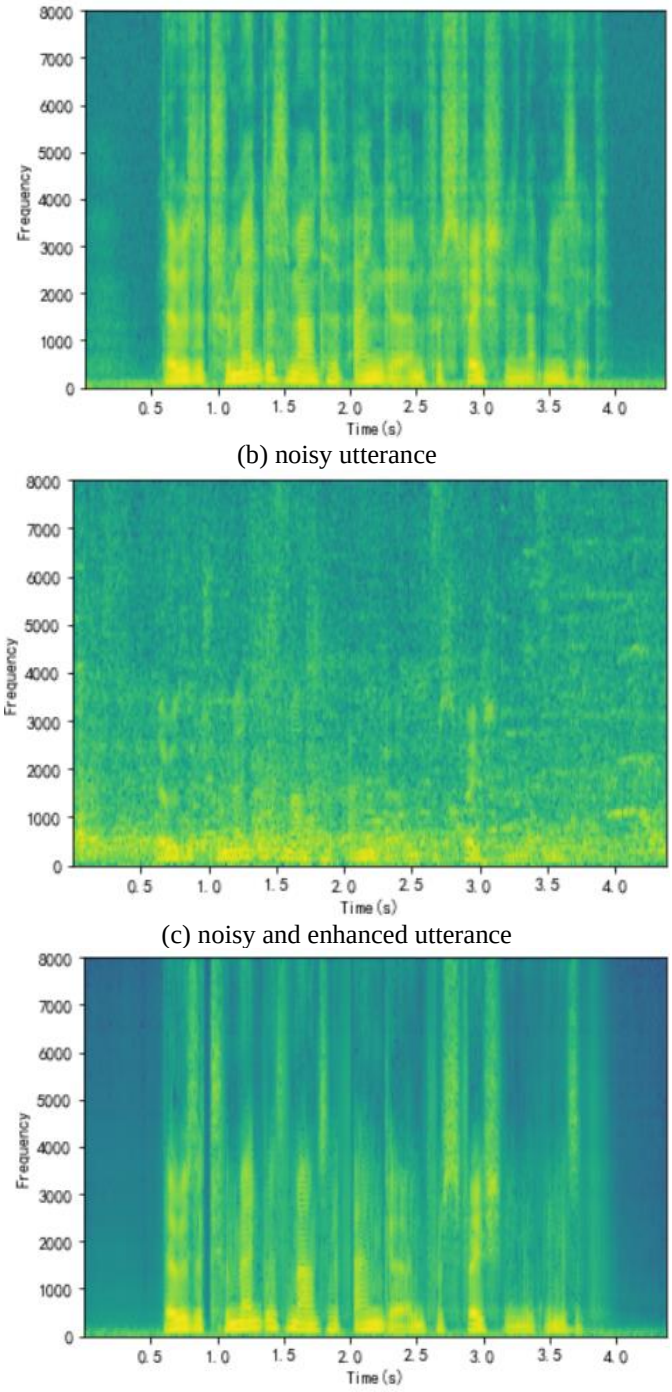


Figure 2. The spectrograms of an utterance at three situations (a) clean (b) noise-corrupted (c) enhanced with the presented method

## VI. Conclusions and Future Works

This study proposes using a monitoring metric loss to determine the best possible DEMUCS network in training. The used monitoring metric loss comprises the original loss and two perceptual metric loss to improve the objective perceptual scores obtained from DEMUCS. Preliminary experimental results confirm the effectiveness of the presented methods in promoting DEMUCS network's SE performance. As a future avenue, we will add other objective metrics, such as scale-invariant signal-to-distortion ratio (SI-SDR), to the monitoring metric loss to see whether DEMUCS can be further enhanced. We will also apply the presented monitoring metric loss to other SE methods to see possible benefits.

## References

[1] Y. Wang, A. Narayanan and D. Wang, "On training targets for supervised speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2014.

[2] P. Karjol, M. Ajay Kumar and P. K. Ghosh, "Speech enhancement using multiple deep neural networks," *in Proc. ICASSP*, 2018.

[3] S. Fu, Y. Tsao, X. Lu and H. Kawai, "Raw waveform-based speech enhancement by fully convolutional networks," *in Proc. APSIPA ASC*, 2017.

[4] A. Dfossez *et al.*, "Music source separation in the waveform domain," *arXiv:1911.13254*, 2019.

[5] A. Defossez, S. Gabriel, and A. Yossi, "Real time speech enhancement in the waveform domain," *in Proc. Interspeech*, 2020

[6] A. W. Rix *et al.*, "Perceptual evaluation of speech quality (PESQ) – a new method for speech quality assessment of telephone networks and codecs," *in Proc. ICASSP*, 2001.

[7] C. H. Taal *et al.*, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Trans. on Audio, Speech, and Language Processing*, 2011.

[8] C. V-Botinhao *et al.*, "Investigating RNN-based speech enhancement methods for noise-robust text-to-speech," *in Proc. SSW*, 2016.

[9] J. Thiemann *et al.*, "Demand: a collection of multi-channel recordings of acoustic noise in diverse environments," *in Proc. ICA*, 2013.