INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Incorporating Embedding Vectors from a Human Mean-Opinion Score Prediction Model for Monaural Speech Enhancement

Khandokar Md. Nayem, Donald S. Williamson

Department of Computer Science, Indiana University, USA
knayem@iu.edu, williads@indiana.edu

Abstract

Objective measures of success, such as the perceptual evaluation of speech quality (PESQ), signal-to-distortion ratio (SDR), and short-time objective intelligibility (STOI), have recently been used to optimize deep-learning based speech enhancement algorithms, in an effort to incorporate perceptual constraints into the learning process. Optimizing with these measures, however, may be sub-optimal, since the objective scores do not always strongly correlate with a listener's evaluation. This motivates the need for approaches that either are optimized with scores that are strongly correlated with human assessments or that use alternative strategies for incorporating perceptual constraints. In this work, we propose an attention-based approach that uses learned speech embedding vectors from a mean-opinion score (MOS) prediction model and a speech enhancement module to jointly enhance noisy speech. Our loss function is jointly optimized with signal approximation and MOS prediction loss terms. We train the model using real-world noisy speech data that has been captured in everyday environments. The results show that our proposed model significantly outperforms other approaches that are optimized with objective measures.

Index Terms: speech enhancement, speech assessment, attention model, deep learning, MOS, speech quality

1. Introduction

Single-channel speech enhancement is a challenging task, but deep-learning based models have shown to be effective in removing unwanted noise and reverberation in certain environments. Deep-learning based speech enhancement approaches traditionally use the mean square error (MSE) between the estimated speech and clean speech signal during training, to optimize performance. This is done due to the computational efficiency of the MSE loss function. However, since MSE is not always a strong indicator of performance, many studies have recently begun to optimize algorithms using other perceptually-inspired objective measures.

Multiple studies use the short-time objective intelligibility (STOI) [1] score to optimize the algorithm and to improve speech intelligibility [2, 3, 4]. Directly optimizing the STOI score is proposed in [3] to minimize the inconsistency between the model optimization criterion and the evaluation criterion for the enhanced speech. The experimental results show that jointly optimizing with STOI and MSE improves speech intelligibility according to objective and subjective measures of success from a listening study. In addition, the word recognition accuracy of the enhanced speech, as assessed by automatic speech recognition (ASR), is improved. Perceptual evaluation of speech quality (PESQ) [5] scores, another popular objective metric, however, have not been increased by optimizing with STOI as reported in [3]. The signal-to-distortion ratio (SDR) [6] has also been used as an objective cost function for optimizing performance [7]. Their results show that optimizing with SDR leads to overall objective quality improvements. Unlike SDR and STOI, PESQ cannot directly be used as an objective function since it is non-differential. Reinforcement learning (RL) techniques such as deep Q-network and policy gradient have thus been employed to solve the non-differentiable problem [4, 8]. In these works, PESQ and the perceptual evaluation methods for audio source separation (PEASS) [9, 10] serve as rewards that are used to optimize model parameters. Meanwhile, a new PESQ-inspired objective function that considers symmetrical and asymmetrical disturbances of speech signals has been developed in [11]. Quality-Net [12], which is a deep neural network (DNN) approach that estimates PESQ scores given a noisy utterance, has also been used as a maximization criteria [13] and as a model selection parameter [14] to enhance speech. Though these papers show that perceptually-inspired speech-quality objective functions can improve performance in certain settings, optimizing with objective measures of success is not always optimal since they do not always strongly correlate with subjective measures [9, 15]. Hence, alternative strategies for incorporating perceptual feedback may be needed.

Subjective evaluation of speech from human listeners remains the gold-standard evaluation approach, since it results in ratings from potential end users. Subjective speech quality evaluations often ask human listeners to either give relative preference scores [16] or assign a numerical rating on the quality of the speech stimuli [17]. Multiple ratings are provided by listeners for each signal, where they are averaged to generate a mean-opinion score (MOS). Recently, various deep-learning approaches have effectively estimated human-assessed MOS [18, 19, 20, 21]. These approaches are promising since they can provide strongly-correlated quality scores for new signals.

Joint learning has been successfully applied in speech enhancement to optimize between estimating speech and other training targets, such as the phase response [22], phoneme class [23], speaker identification [24], and speech recognition [25]. In a similar manner, we propose to leverage the benefits of speech-quality estimation for the speech enhancement task using joint learning. In particular, we propose an attention-based speech enhancement model that uses the embedding vector from a MOS prediction model to produce speech with better perceptual quality. The MOS estimator generates encoded embedding vectors from the input noisy speech. Our speech enhancement attention model is conditioned on that embedding vector and enhances the noisy speech using a separate encoder-decoder framework. The embedding vector extracts perceptually useful features that are important for human-based assessment. The speech enhancement model will leverage these features, which should help produce better quality speech according to human evaluation. Our proposed model jointly updates

http://dx.doi.org/10.21437/Interspeech.2021-1844

both the MOS-prediction and speech-enhancement models during training, using speech enhancement and MOS prediction loss terms.

The rest of the paper is organized as follows. In section 2, we introduce the speech quality assessment model and the proposed speech enhancement model. We describe our dataset, experimental setup and evaluate our proposed joint learning approach in section 3. We conclude our work in section 4.

2. Proposed Approach

Let's define s_t as the clean speech signal and n_t as the noise at time t. The mixture of clean speech and noise is denoted as $m_t = s_t + n_t$. We aim to extract the speech from the mixture by removing the unwanted noise. The short-time Fourier transform (STFT) is first used to convert the time-domain mixture signal into a time-frequency (T-F) domain signal, $M_{t,f}$, that is defined at time t and frequency f. The complex-valued STFT matrix, $\boldsymbol{M}$, can be written as $\boldsymbol{M} = |\boldsymbol{M}|e^{i\boldsymbol{\theta}^M}$ with magnitude $|\boldsymbol{M}| \in \Re_+^{T\times F}$ and phase $\boldsymbol{\theta}^M \in \Re^{T\times F}$. The enhancement of the noisy speech magnitude directly produces an estimated clean magnitude response $|\hat{\boldsymbol{S}}|$, using an enhancement function $\mathcal{F}$ such that $|\hat{\boldsymbol{S}}| = \mathcal{F}(|\boldsymbol{M}|)$. The enhancement function is modeled with a deep neural network. This estimated magnitude response is then combined with the noisy phase, $\boldsymbol{\theta}^M$, where the inverse STFT can subsequently be used to produce an enhanced speech signal in the time domain, $\hat{s}_t$.

A depiction of our speech-enhancement model is shown in Figure 1. The model consists of a MOS prediction model (shown left - red box) and a speech enhancement model (shown right - blue box). We next will describe each of these sub-modules.

2.1. Speech quality assessment model

We adapt the data-driven MOS prediction model from [26] to estimate MOS from the noisy speech signals. This model has been developed with real-world data and it has been shown to outperform comparison approaches [12, 18, 27], according to multiple metrics. The MOS prediction model consists of an attention-based encoder-decoder structure that uses stacked pyramid bi-directional long-short term memory (pBLSTM) [28] networks in the encoder. We denote this MOS prediction model as Pyramid-MOS (PMOS). A pBLSTM architecture gives the advantages of processing sequences at multiple time resolutions, which effectively captures short- and long-term dependencies. Speech has spectral and temporal dependencies over short and long durations, and a multi-resolution framework is effective in learning these complex relations.

The input to the network is a one-time frame of a noisy-speech mixture $|\boldsymbol{M}_t|$. In a pyramid structure, the lower layer outputs from Υ consecutive time frames are concatenated and used as inputs to the next pBLSTM layer, along with the recurrent hidden states from the previous time step. The output of a pBLSTM node is an embedding vector, h_t^l, that is as defined below:

$$h_t^l = pBLSTM\left(h_{t-1}^l, \left[h_{\Upsilon\times t-\Upsilon+1}^{l-1}, h_{\Upsilon\times t}^{l-1}\right]\right) \quad (1)$$

where Υ is the reduction factor between successive pBLSTM layers and l is the layer number. A pyramid-BLSTM structure reduces the time resolution from the input speech to the final latent representation $\boldsymbol{H}$. This compressed vector accumulates the useful features for measuring speech perceptual quality

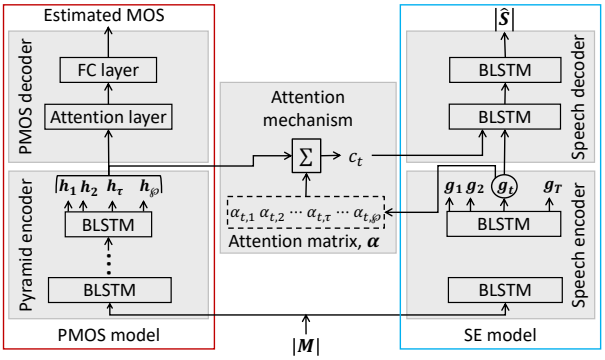

Figure 1: *A depiction of our speech-enhancement model that consists of a MOS-prediction model denoted as PMOS (left side -red box), and a speech-enhancement (SE) model (right side - blue box). An attention mechanism connects the two models.*

that resides in a range of time-frames and ignores the least important features. The encoder output is generated by concatenating the hidden states of the last pBLSTM layer into vector $\boldsymbol{H} = \{\boldsymbol{h}_1, \boldsymbol{h}_2, \cdots, \boldsymbol{h}_\tau, \cdots, \boldsymbol{h}_\wp\}$, where $\wp$ is the total number of final embedding vectors with index τ.

The decoder of the PMOS model is implemented as an attention layer followed by a fully-connected (FC) layer and it outputs an estimated MOS of the input speech. Attention models learn key attributes of a latent sequence, since adjacent time frames can provide important information, which is particularly necessary for our task. The self-attention mechanism [29] uses the pyramid encoder output at the i-th and k-th time steps to compute the attention weights, $\alpha_{i,k}^{PMOS}$. Attention weights are used to compute context vector c_i^{PMOS} using the below equations:

$$\alpha_{i,k}^{PMOS} = \frac{\exp\left(\boldsymbol{h}_i^\top \boldsymbol{Q}\boldsymbol{h}_k\right)}{\sum_{i=1}^{\wp}\exp\left(\boldsymbol{h}_i^\top \boldsymbol{Q}\boldsymbol{h}_k\right)} \quad (2)$$

$$c_i^{PMOS} = \sum_{k=1}^{\wp} \alpha_{i,k}^{PMOS} \cdot \boldsymbol{h}_k \quad (3)$$

Here $\boldsymbol{Q}$ is the PMOS attention weight matrix. The context vector is provided to a fully-connected (FC) layer to estimate the MOS of the noisy speech signal. Note that the pyramid structure of the encoder results in a shorter sequence of latent representations than the original input sequence, and it leads to fewer encoding states for attention calculation at the decoding stage. Therefore, strictly $\wp < T$, and in our case $\wp = \lceil T/\Upsilon^L \rceil$, where L is the number of pBLSTM layers.

2.2. Attention-based speech enhancement

Our proposed speech-enhancement (SE) model also follows an encoder-decoder structure, and it is shown in Figure 1 at the right (blue box). The SE encoder takes a single time-frame of a noisy-speech mixture, $|\boldsymbol{M}_t|$, as input and multiple BLSTM layers, are stacked together to create a hidden representation of the frame, $\boldsymbol{g}_t$. An attention mechanism [30] is applied using the mixture encoding from the SE model, $\boldsymbol{G} = \{\boldsymbol{g}_1, \boldsymbol{g}_2, \cdots, \boldsymbol{g}_T\}$, and the PMOS encoding, $\boldsymbol{H}$, from the MOS prediction model. This allows the SE model to exploit the MOS estimator's encoding and utilize the important perceptual feature embedding that correlates with human assessment. Since PMOS yields encoding vector $\boldsymbol{H}$, which has a smaller time resolution than the encoding from the SE encoder, we compute a score for each embedding vector $\boldsymbol{h}_\tau^L$ using a learnable weight matrix, $\boldsymbol{W}$. Then

the attention weights for the SE model, $\alpha_{t,\tau}$, are obtained using a softmax operation over the scores of all $\boldsymbol{h}_\tau^L$. Now, the PMOS encoding is summarized in a context vector $\boldsymbol{c}_t$ for each mixture frame $\boldsymbol{g}_t$. Prior to computing $\boldsymbol{c}_t$, $\boldsymbol{h}_\tau^L$ passes through a linear layer ℓ, so that we learn a different representation for the SE task. The computations are below:

$$score_{t,\tau} = \boldsymbol{g}_t^\top \boldsymbol{W} \boldsymbol{h}_\tau \tag{4}$$

$$\alpha_{t,\tau} = \frac{\exp\left(score_{t,\tau}\right)}{\sum_{t=1}^{T} \exp\left(score_{t,\tau}\right)} \tag{5}$$

$$\boldsymbol{c}_t = \sum_{\tau=1}^{\wp} \alpha_{t,\varepsilon} \cdot \ell(\boldsymbol{h}_\tau) \tag{6}$$

Since we are learning two targets MOS and enhanced speech simultaneously, the unified model will learn different representations for these tasks. Thus both PMOS and SE models will learn their corresponding targets with perceptual feature sharing. Then, the context vector and SE-model embedding vector are concatenated (e.g., $[\boldsymbol{c}_t, \boldsymbol{g}_t]$) and passed to the decoder module. The SE-decoder module follows the network structure from [23]. It consists of a linear layer with a $tanh(\cdot)$ activation function, two BLSTM layers, and a linear layer with ReLU activation. It outputs the estimated enhanced speech $|\hat{\boldsymbol{S}}|$.

2.3. Joint-learning objective function

Our joint-learning objective function uses a weighted average of a time-domain signal-approximation loss $\mathcal{L}_{sa}$ (from the SE model), the MSE of the magnitude spectrum $\mathcal{L}_{mse}$ (from the SE model) and the MSE of the MOS estimation $\mathcal{L}_{mos}$ (from the PMOS model). We compute the signal-approximation loss from the time-domain signal difference between the reference speech s and enhanced speech $\hat{s}$. The overall loss function of our network is defined as below, using hyper-parameters λ_1 and λ_2 that control the impact of individual loss terms:

$$\mathcal{L} = \lambda_1 \left[\lambda_2 \mathcal{L}_{mse} + (1-\lambda_2)\mathcal{L}_{sa}\right] + (1-\lambda_1)\mathcal{L}_{mos} \tag{7}$$

We first train the PMOS model using $\mathcal{L}_{mos}$ (e.g. $\lambda_1 = 0$), then we train the SE model using $\lambda_1 = 1$, while running the PMOS model in inference mode (e.g. it is held fixed). This is done to ensure that the trained PMOS model effectively encodes the key features in the embedding vector that are important to perceptual speech quality. Finally, we train both the models jointly using $\mathcal{L}$ to further reduce any correctional differences between the true MOS and estimated MOS in the PMOS model, and to increase the perceptual quality of the enhanced speech.

3. Experiments and Results

3.1. Dataset

For training and testing, we use the COnversational Speech In Noisy Environments (COSINE) [31] and Voices Obscured in Complex Environmental Settings (VOiCES) [32] corpora. The COSINE corpus contains 150 hours of audio recordings that are captured using 7-channel wearable microphones, with multiparty conversations in a variety of noisy environments (e.g., street, cafeteria, bus, wind noise, etc). Audio captured by the close-talking microphone is used as the clean reference, whereas audio from the shoulder and chest microphones are considered noisy signals with significant amounts of background noise. The VOiCES corpus records audio using 12 microphones placed throughout two rooms of different sizes.

Table 1: *Performance comparison with MOS prediction models comparing against the ground truth MOS obtained from human subjects. Best results are shown in* ***bold****.*

	MAE	RMSE	PCC (γ)	SRCC (ρ)
NISQA [27]	0.62	0.7	0.71	0.79
PMOS [26]	0.51	0.57	0.88	0.88
SE+PMOS	**0.45**	**0.52**	**0.9**	**0.91**

Different background noises are played separately in conjunction with foreground clean speech, so the signals contain noise and reverberation. Foreground speech is used as the reference clean signal, and the audio from two microphones is used as the reverberant-noisy speech. The approximated speech-to-reverberation ratios (SRRs) of the VOiCES signals range from -4.9 to 4.3 dB. In COSINE, the approximated signal-to-noise ratios (SNRs) range from -10.1 to 11.4 dB. The MOS data was captured from the listening study that is outlined in [26], which contains MOS quality ratings for 18,000 COSINE signals and 18,000 VOiCES signals. In total, 45 hours of speech signals are generated and 180k subjective human judgments are collected.

Both speech corpora consist of 16-bit single-channel files sampled at 16 kHz. For MOS prediction, the input speech signals are segmented into 40 ms length frames, with 10 ms overlap. A FFT length of 512 samples and a Hanning window are used to compute the spectrogram. Mean and variance normalization are applied to the input feature vector. Noisy or reverberant stimuli of each dataset are divided into training (70%), validation (10%), and testing (20%) sets, and trained separately.

3.2. Network architecture

The PMOS encoder uses $L = 3$ pBLSTM layers (with 128, 64 and 32 nodes in each direction, respectively) on top of a BLSTM layer that has 256 nodes. As in [26, 28], the reduction factor $\Upsilon = 2$ is adopted here. Therefore, the final latent representation $\boldsymbol{h}_\tau$ is reduced in the time resolution by a factor of $\Upsilon^3 = 8$. In the PMOS decoder, the context vector is passed to a FC layer with 32 units. The model is optimized using Adam optimization with convergence determined by a validation set.

Our proposed SE model uses a LSTM based encoder-decoder architecture, where the encoder consists of 2 BLSTM layers. The decoder has a linear layer with $tanh$ activation, followed by a 2-layer BLSTM and a linear layer with ReLU activation [23, 34]. Each LSTM layer contains 200 nodes and the linear layer has 321 nodes. The input feature vector is the magnitude of the mixture spectrogram computed using a hamming window with 50% overlap after normalization. Adam optimization [35] is applied and the learning rate is 0.0001. An early stopping method is used if the performance on the validation does not decrease in 200 consecutive epochs.

3.3. Results

We use four metrics to evaluate MOS-estimation performance; mean absolute error (MAE), epsilon insensitive root mean squared error (RMSE) [36], Pearson's correlation coefficient γ (PCC), and Spearman's rank correlation coefficient ρ (SRCC). We evaluate our proposed model in two stages. In the first stage, we train our MOS estimation model (PMOS) [26]. Then we freeze the PMOS model and train the speech enhancement (SE) model using learned PMOS embeddings. Finally, we perform joint learning of the PMOS and SE models, and this is denoted as SE+PMOS. In Table 1, we compare MOS prediction performance with NISQA [27], which is modified to estimate MOS

Table 2: *Average results of the speech enhancement models in different performance metrics. Best results are shown in* ***bold****.*

		COSINE				VOiCES			
	loss func.	PESQ	SI-SDR	ESTOI	MOS-LQO	PESQ	SI-SDR	ESTOI	MOS-LQO
Mixture	-	1.46	0.53	0.62	4.04	1.26	-1.3	0.48	2.74
SE	$\mathcal{L}_{mse}$	2.68	2.8	0.8	3.2	2.3	1.2	0.69	3.5
	$\mathcal{L}_{mos}$ [13]	2.8	3.8	0.82	4.2	2.37	1.66	0.74	5.3
	$\mathcal{L}_{mse}, \mathcal{L}_{sa}$	2.72	3.1	0.82	4	2.35	1.6	0.7	3.8
	$\mathcal{L}_{sa}, \mathcal{L}_{mos}$	2.89	4.1	0.85	4.4	2.42	1.72	0.77	5.7
	$\mathcal{L}_{sdr}$ [7]	2.7	4.5	0.82	4	2.32	2.01	0.72	4.5
SE+PMOS	$\mathcal{L}_{mse}$	3.1	4	0.85	4.2	2.48	1.8	0.8	6
	$\mathcal{L}_{mse}, \mathcal{L}_{sa}$	3.19	**4.6**	0.93	4.8	2.54	**2.08**	0.86	6.3
	$\mathcal{L}_{mse}, \mathcal{L}_{sa}, \mathcal{L}_{mos}$	3.19	4.5	0.92	**5.1**	2.53	2.06	0.84	**6.5**
MetricGAN [33]	$\mathcal{L}_{pesq}$	**3.28**	4.4	0.9	5	**2.67**	2.01	0.83	6.1
	$\mathcal{L}_{stoi}$	3.19	4.3	**0.94**	4.8	2.5	2	**0.87**	5.8
SSEMS [14]	$\mathcal{L}_{qnet}(\phi = 0dB)$	2.85	2.9	0.83	3	2.4	1.8	0.7	2.8

using a convolutional neural network (CNN) and BLSTM architecture. Results show that the proposed SE+PMOS clearly outperforms other MOS prediction models according to all metrics. The absolute correlation error minimizes by 0.6 compared to the original PMOS [26] approach. This justifies the use of our PMOS model, but also points to the benefits of joint learning.

In terms of speech enhancement, we compare against a speech enhancement model without attention mechanism [37] and denote this baseline model as SE. Different loss functions are used to optimize this model, including MSE, MSE plus signal approximation, MOS, signal approximation with MOS, and SDR. We compare our proposed approach against a generative adversarial network (GAN) that individually optimizes with PESQ ($\mathcal{L}_{pesq}$) and STOI ($\mathcal{L}_{stoi}$) [33]. We denote this model as MetricGAN. We compare against an ensemble-based specialized speech enhancement model selection (SSEMS) [14] approach that uses Quality-Net [12], which estimates PESQ scores, as their objective function. They form an ensemble structure with several enhancement models, where each model is trained using audio at specific SNRs. During inference, they choose the output that has the highest objective PESQ score. We choose the SNR threshold $\phi = 0dB$ for balance training. All models are trained using the experimental setup that is previously mentioned. We modify the comparison models using the code provided by the original authors. Speech enhancement performance is assessed with PESQ [5], scale-invariant SDR (SI-SDR) [6, 38], and extended STOI (ESTOI) [1, 39]. Additionally, we measure predicted MOS score of enhanced speech, using our PMOS model, since we aim to improve human-assessed speech quality. We denote this metric as MOS-LQO.

Table 2 shows the average results of the different enhancement models, according to each of the performance metrics. As the scores of the unprocessed mixture show, the reverberant VOiCES corpus is much more challenging than the noisy COSINE corpus. For the baseline SE model, we experiment with 5 different combinations of loss functions. With the MSE loss, $\mathcal{L}_{mse}$, we see improvements in objective scores, except MOS-LQO for the COSINE data. With $\mathcal{L}_{mos}$ as the only objective criteria as proposed in [13], MOS-LQO improves by 1.4 overall compared with SE with $\mathcal{L}_{mse}$. Then we separately combine the signal approximation loss with MSE loss and MOS loss. In terms of PESQ, we gain on average approximately 0.35 and 0.5 compared to the models that use only the MSE loss and MOS loss, respectively. This suggests that $\mathcal{L}_{mse}$ and $\mathcal{L}_{sa}$ maximize the overall objective intelligibility, whereas $\mathcal{L}_{mos}$ focuses more on perceptual speech quality. Note that in all these $\mathcal{L}_{mos}$ calculation, we use a separately trained PMOS model (e.g., no joint learning). Lastly, we apply the SDR loss function as proposed in [7]. We observe an average gain of 0.35 in SI-SDR, however, it yields a poor score according to other metrics, especially a 0.8 loss in MOS compared to SE with $\mathcal{L}_{mse}, \mathcal{L}_{sa}$ loss terms. We calculate the performance of our proposed model using three combinations of loss functions. Using $\mathcal{L}_{mse}$ and $\mathcal{L}_{sa}$, we achieve the highest SI-SDR scores for both corpora, though these results are nearly identical to the model trained with all three loss terms (e.g., $\mathcal{L}$ (eq:7)). Using all three loss terms, the MOS-LQO score is 5.1 and 6.5 for noisy and reverberant environments, respectively, which are the highest results for this metric. MetricGAN optimizes PESQ or STOI, hence, it outperforms other models in terms of PESQ and ESTOI separately, though the scores for our two- and three-term SE+PMOS approaches are only slightly lower even though PESQ and STOI are not considered during training. SSEMS approach yields the lowest scores across all metrics. It is important to note that proposed approach performs best according to SI-SDR, where this metric is not used by any of the approaches during optimization.

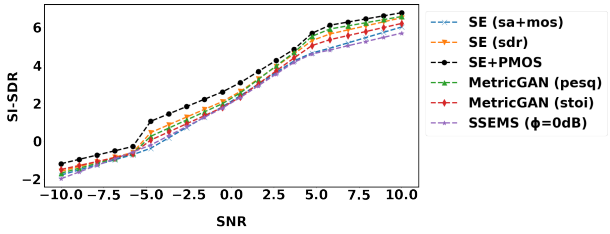

Figure 2: *Average SI-SDR performance of speech enhancement models on test speech in different SNRs.*

Figure 2 shows the SI-SDR performance of the different models as a function of SNR. Our SE+PMOS approach outperforms all the other models at each SNR. Similar results are observed for the other metrics, except at low SNRs for PESQ and ESTOI, where MetricGAN performs slightly better.

4. Conclusion

Our proposed speech enhancement model utilizes a speech quality MOS assessment metric in a joint learning manner and the results show that it outperforms other models in both noisy and reverberant environments. It shows that perceptually-relevant embeddings are useful for speech enhancement. However, we evaluate our model's subjective score using a MOS-estimation model. Additionally, our assessment model provides utterance-level feedback, which may be sub-optimal since the model's embeddings are calculated at the frame level. These will be addressed in future work.

5. References

[1] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE/ACM TASLP*, vol. 19, pp. 2125–2136, 2011.

[2] H. Zhang, X. Zhang, and G. Gao, "Training supervised speech separation system to improve STOI and PESQ directly," in *Proc. ICASSP*. IEEE, 2018, pp. 5374–5378.

[3] S. W. Fu, T. W. Wang, Y. Tsao, X. Lu, and H. Kawai, "End-to-end waveform utterance enhancement for direct evaluation metrics optimization by fully convolutional neural networks," *IEEE/ACM TASLP*, vol. 26, pp. 1570–1584, 2018.

[4] Y. Koizumi, K. Niwa, Y. Hioka, K. Kobayashi, and Y. Haneda, "Dnn-based source enhancement to increase objective sound quality assessment score," *IEEE/ACM TASLP*, vol. 26, pp. 1780–1792, 2018.

[5] 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 *Proc. ICASSP*, vol. 2, 2001, pp. 749–752.

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

[7] M. Kawanaka, Y. Koizumi, R. Miyazaki, and K. Yatabe, "Stable training of dnn for speech enhancement based on perceptually-motivated black-box cost function," in *Proc. ICASSP*. IEEE, 2020, pp. 7524–7528.

[8] Y. Koizumi, K. Niwa, Y. Hioka, K. Kobayashi, and Y. Haneda, "Dnn-based source enhancement self-optimized by reinforcement learning using sound quality measurements," in *Proc. ICASSP*. IEEE, 2017, pp. 81–85.

[9] V. Emiya, E. Vincent, N. Harlander, and V. Hohmann, "Subjective and objective quality assessment of audio source separation," *IEEE TASLP*, vol. 19, pp. 2046–2057, 2011.

[10] E. Vincent, "Improved perceptual metrics for the evaluation of audio source separation," in *International Conference on Latent Variable Analysis and Signal Separation*. Springer, 2012, pp. 430–437.

[11] J. M. Martín-Doñas, A. M. Gomez, J. A. Gonzalez, and A. M. Peinado, "A deep learning loss function based on the perceptual evaluation of the speech quality," *IEEE Signal processing letters*, vol. 25, pp. 1680–1684, 2018.

[12] S.-W. Fu, Y. Tsao, H.-T. Hwang, and H.-M. Wang, "Quality-net: An end-to-end non-intrusive speech quality assessment model based on blstm," in *Proc. ICASSP*, 2018.

[13] S.-W. Fu, C.-F. Liao, and Y. Tsao, "Learning with learned loss function: Speech enhancement with quality-net to improve perceptual evaluation of speech quality," *IEEE Signal Processing Letters*, vol. 27, pp. 26–30, 2019.

[14] R. E. Zezario, S.-W. Fu, X. Lu, H.-M. Wang, and Y. Tsao, "Specialized speech enhancement model selection based on learned non-intrusive quality assessment metric." in *Proc. Interspeech*, 2019, pp. 3168–3172.

[15] A. W. Rix, J. G. Beerends, D.-S. Kim, P. Kroon, and O. Ghitza, "Objective assessment of speech and audio quality—technology and applications," *IEEE TASLP*, vol. 14, pp. 1890–1901, 2006.

[16] S. R. Quackenbush, T. P. Barnwell III, and M. A. Clements, *Objective Measures Of Speech Quality*, 1st ed. Prentice Hall, 1988.

[17] L. Malfait, J. Berger, and M. Kastner, "P. 563—the itu-t standard for single-ended speech quality assessment," *IEEE TASLP*, vol. 14, pp. 1924–1934, 2006.

[18] A. R. Avila, H. Gamper, C. Reddy, R. Cutler, I. Tashev, and J. Gehrke, "Non-intrusive speech quality assessment using neural networks," in *Proc. ICASSP*. IEEE, 2019, pp. 631–635.

[19] B. Patton, Y. Agiomyrgiannakis, M. Terry, K. Wilson, R. A. Saurous, and D. Sculley, "Automos: Learning a non-intrusive assessor of naturalness-of-speech," *NIPS Workshop*, 2016.

[20] C.-C. Lo, S.-W. Fu, W.-C. Huang, X. Wang, J. Yamagishi, Y. Tsao, and H.-M. Wang, "MOSNet: Deep learning based objective assessment for voice conversion," in *Proc. Proc.*, 2019.

[21] X. Dong and D. S. Williamson, "An attention enhanced multi-task model for objective speech assessment in real-world environments," in *Proc. ICASSP*. IEEE, 2020, pp. 911–915.

[22] J. Lee and H.-G. Kang, "A joint learning algorithm for complex-valued tf masks in deep learning-based single-channel speech enhancement systems," *IEEE/ACM TASLP*, vol. 27, pp. 1098–1108, 2019.

[23] K. Schulze-Forster, C. S. Doire, G. Richard, and R. Badeau, "Joint phoneme alignment and text-informed speech separation on highly corrupted speech," in *Proc. ICASSP*. IEEE, 2020, pp. 7274–7278.

[24] X. Ji, M. Yu, C. Zhang, D. Su, T. Yu, X. Liu, and D. Yu, "Speaker-aware target speaker enhancement by jointly learning with speaker embedding extraction," in *Proc. ICASSP*. IEEE, 2020, pp. 7294–7298.

[25] C. Donahue, B. Li, and R. Prabhavalkar, "Exploring speech enhancement with generative adversarial networks for robust speech recognition," in *Proc. ICASSP*. IEEE, 2018, pp. 5024–5028.

[26] X. Dong and D. S. Williamson, "A Pyramid Recurrent Network for Predicting Crowdsourced Speech-Quality Ratings of Real-World Signals," in *Proc. Interspeech*, 2020, pp. 4631–4635.

[27] G. Mittag and S. Möller, "Non-intrusive speech quality assessment for super-wideband speech communication networks," in *Proc. ICASSP*. IEEE, 2019, pp. 7125–7129.

[28] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *Proc. ICASSP*. IEEE, 2016, pp. 4960–4964.

[29] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, pp. 5998–6008, 2017.

[30] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," in *Proc. Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2015, pp. 1412–1421.

[31] A. Stupakov, E. Hanusa, J. Bilmes, and D. Fox, "Cosine-a corpus of multi-party conversational speech in noisy environments," in *Proc. ICASSP*. IEEE, 2009, pp. 4153–4156.

[32] C. Richey, M. A. Barrios, Z. Armstrong, C. Bartels, H. Franco, M. Graciarena, A. Lawson, M. K. Nandwana, A. Stauffer, J. van Hout *et al.*, "Voices obscured in complex environmental settings (voices) corpus," *arXiv preprint arXiv:1804.05053*, 2018.

[33] S.-W. Fu, C.-F. Liao, Y. Tsao, and S.-D. Lin, "Metricgan: Generative adversarial networks based black-box metric scores optimization for speech enhancement," in *Proc. ICML*, 2019.

[34] K. Schulze-Forster, C. Doire, G. Richard, and R. Badeau, "Weakly informed audio source separation," in *IEEE WASPAA*. IEEE, 2019, pp. 273–277.

[35] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR*, Y. Bengio and Y. LeCun, Eds., 2015.

[36] I. Rec, "P. 1401, methods, metrics and procedures for statistical evaluation, qualification and comparison of objective quality prediction models," *International Telecommunication Union, Geneva, Switzerland*, 2012.

[37] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Proc. ICASSP*. IEEE, 2013, pp. 6645–6649.

[38] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR–half-baked or well done?" in *Proc. ICASSP*, 2019, pp. 626–630.

[39] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM TASLP*, vol. 24, pp. 2009–2022, 2016.