PERCEPTUALLY GUIDED SPEECH ENHANCEMENT USING DEEP NEURAL NETWORKS

Yan Zhao[1] *Buye Xu*[2] *Ritwik Giri*[2] *Tao Zhang*[2]

[1]Department of Computer Science and Engineering, The Ohio State University, USA
[2]Starkey Hearing Technologies, USA

zhao.836@osu.edu, {buye_xu, ritwik_giri, tao_zhang}@starkey.com

ABSTRACT

Human listeners often have difficulties understanding speech in the presence of background noise in the real world. Recently, supervised learning based speech enhancement approaches have achieved substantial success, and show significant improvements over the conventional approaches. However, existing supervised learning based approaches often try to minimize the mean squared error between the enhanced output and the pre-defined training target (e.g., the log power spectrum of clean speech), even though the purpose of such speech enhancement is to improve speech understanding in noise. In this paper, we propose a new deep neural networks based enhancement approach by incorporating a speech perception model into the loss function. Specifically, we use the short-time objective intelligibility metric in the loss in addition to the mean squared error. Optimizing the proposed perceptually guided loss is expected to improve speech intelligibility further. Systematic evaluations show that our proposed approach is able to improve speech intelligibility in a wide range of signal-to-noise ratios and noise types while maintaining speech quality.

Index Terms— ideal ratio mask, denoising, speech intelligibility, STOI, deep neural networks

1. INTRODUCTION

In real-world environments, speech is inevitably corrupted by background noise coming from various sound sources like other speakers, machines and so forth. These distortions degrade both speech intelligibility and quality, especially when the signal-to-noise ratio (SNR) is at low level. For both normal hearing (NH) and hearing impaired (HI) listeners, understanding noisy speech usually becomes very challenging. This is detrimental to effective communication among people. On the other hand, many speech-related applications, including automatic speech recognition (ASR) and speaker identification (SID), perform poor under adverse noisy conditions [1, 2].

Enhancing speech in noise has attracted considerable research efforts in the past decades. In recent years, many deep learning based supervised speech enhancement approaches have been proposed and substantial performance improvements have achieved over conventional signal processing based approaches. The key idea is to formulate the denoising problem as a supervised learning task, and then employ the deep learning techniques to solve it. Xu *et al.* [3] propose to utilize deep neural networks (DNN) to learn a non-linear mapping function from the log power spectrum of noisy speech to that of the corresponding clean speech. Instead of performing direct mapping, Wang *et al.* [4] employ a set of complementary features extracted from corrupted speech to estimate the ideal ratio mask (IRM), and then apply the predicted ratio mask to the time-frequency (T-F) representation of noisy speech to obtain the enhanced speech. Considering that the estimation of T-F mask is an intermediate result which does not directly lead to the actual enhancement objective, Weninger *et al.* [5] propose a signal approximation loss function. Optimizing the new loss function by using the long short-term memory deep recurrent neural networks (LSTM-DRNN) improves the performance of T-F masking approach further. Erdogan *et al.* [6] develop a phase-sensitive mask, which incorporates the phase difference between noisy speech and clean speech, resulting in good performance in terms of signal-to-distortion ratio (SDR). Wang and Wang [7] propose to optimize a loss function defined in the time domain, where the enhanced time-domain signal is reconstructed by using noisy phase during training. It has been shown that computing the loss in the time domain is equivalent to the phase-sensitive masking approach [8]. Zhao *et al.* [9] extend Wang and Wang's time domain reconstruction approach by using clean phase during training to obtain a better estimate of magnitude. In order to jointly enhance the magnitude spectrum and phase spectrum, Williamson *et al*. [10] propose the complex ideal ratio mask (cIRM), and perform T-F masking in the complex domain. Since the noisy phase is also enhanced, better speech quality is reported.

Significant improvements over traditional speech enhancement approaches have been reported in previous studies. Existing supervised enhancement approaches are trained to minimize the mean squared error (MSE) between the output and the corresponding training target (e.g., log power spectrum of clean signal, or IRM). In the ideal case, when the MSE is minimized to zero, the processed signal is restored to the ideal target, and thus the perceptual aspects (i.e. sound quality and intelligibility) would be optimized. However, in practice the MSE cannot be reduced to zero, and the residual can be high, especially when the SNR of the input signal is low. Although related, the MSE criterion does not directly reflect the perceived speech quality and intelligibility. In other words, from the perspective of human listeners, the MSE is not the optimal objective to optimize. It is desired to leverage the domain knowledge of speech perception in the loss function. This paper attempts to directly incorporate the short-time objective intelligibility measure (STOI) [11] in a supervised speech enhancement approach to optimize for speech intelligibility. The popular STOI metric has shown high correlation with speech intelligibility.

One work that is closely related to ours is proposed by Koizumi *et al.* [12]. In their study, the perceptual metrics are introduced to optimize the speech enhancement algorithm. Specifically, the perceptual evaluation of speech quality (PESQ) [13] and perceptual evaluation methods for audio source separation (PEASS) [14] are used to design a time varying reward. A set of mask templates are defined as actions. Then the DNN-based speech enhancement algorithm is

This work was conducted when Yan Zhao did a signal processing research internship at Starkey.

optimized by utilizing reinforcement learning (RL) with the previously defined reward and actions. Different from their approach, we directly incorporate a speech intelligibility metric into the loss function and optimize it by supervised learning.

The rest of the paper is organized as follows. In next section, we describe the proposed approach in details. The experimental setup and evaluation results are presented in Section 3 and Section 4, respectively. Finally, we conclude this paper in Section 5.

2. ALGORITHM DESCRIPTION

In this section, we will introduce the proposed perceptually guided speech enhancement approach, including the modified STOI computation and the loss function.

2.1. Modified STOI computation

The original STOI metric is described in details in [11]. It is calculated in the short-term one-third-octave-band domain with a window length of 384 ms. However, the supervised speech enhancement approach in this study is performed in the short-time Fourier transform (STFT) domain with a 32 ms Hanning window and a 16 ms window shift. Assuming a 16 kHz sampling rate, for each time frame, a 512-point fast Fourier transformation (FFT) is applied, resulting in 257 frequency bins. In order to comply with the STOI calculation, the frequency bins are grouped to form one-third octave bands. Specifically, let $X(m,f)$, $Y(m,f)$ denote the STFT representation of the clean reference signal and the enhanced signal, respectively, at time frame m and frequency channel f. Corresponding frequency bins are then combined to 15 one-third octave bands, where the center frequency is set from 150 Hz to around 4.3 kHz. Then, we have the new T-F representations as follows,

$$\begin{aligned} X_j(m) &= \sqrt{\sum_{f=f_1(j)}^{f_2(j)-1} \|X(m,f)\|_2^2} \\ Y_j(m) &= \sqrt{\sum_{f=f_1(j)}^{f_2(j)-1} \|Y(m,f)\|_2^2} \end{aligned} \tag{1}$$

where j is the index of the one-third octave band; f_1 and f_2 are the edges of the one-third octave bands; $\|\cdot\|_2$ denotes the L_2 norm. Then, the short-term temporal envelope of the clean speech and the enhanced speech can be denoted by the following two vectors,

$$\begin{aligned} \mathbf{x}_{m,j} &= [X_j(m), X_j(m+1), ..., X_j(m+N-1)]^T \\ \mathbf{y}_{m,j} &= [Y_j(m), Y_j(m+1), ..., Y_j(m+N-1)]^T \end{aligned} \tag{2}$$

where N is set to 24 corresponding to the 384 ms analysis window length. According to the original STOI computation, the short-term temporal envelope of the enhanced speech is normalized and clipped by using the following equation,

$$\bar{\mathbf{y}}_{m,j}(i) = min(\frac{\|\mathbf{x}_{m,j}\|_2}{\|\mathbf{y}_{m,j}\|_2}\mathbf{y}_{m,j}(i), (1+10^{-\beta/20})\mathbf{x}_{m,j}(i)) \tag{3}$$

where $i = 1, 2, ..., N$; β controls the lower bound of SDR, which is set to -15 in our study following the original STOI implementation.

The correlation coefficient between the vectors $\mathbf{x}_{m,j}$ and $\bar{\mathbf{y}}_{m,j}$ is defined as the intermediate speech intelligibility measure, namely,

$$d_{m,j} = \frac{(\mathbf{x}_{m,j} - \mu_{\mathbf{x}_{m,j}})^T(\bar{\mathbf{y}}_{m,j} - \mu_{\bar{\mathbf{y}}_{m,j}})}{\|\mathbf{x}_{m,j} - \mu_{\mathbf{x}_{m,j}}\|_2 \|\bar{\mathbf{y}}_{m,j} - \mu_{\bar{\mathbf{y}}_{m,j}}\|_2} \tag{4}$$

where $\mu_{(\cdot)}$ denotes the sample mean of the vector.

The speech intelligibility at time frame m can be calculated by taking average over all one-third octave bands. We define a modified STOI function at time frame m as follows,

$$d_m = f(\boldsymbol{X}_m^{24}, \boldsymbol{Y}_m^{24}) = \frac{1}{J}\sum_j d_{m,j} \tag{5}$$

where $\boldsymbol{X}_m^{24}$ and $\boldsymbol{Y}_m^{24}$ denote the 24-frame magnitude spectrum starting from the time frame m of the clean reference speech and the corresponding enhanced speech, respectively; J denotes the total number of the one-third octave bands. It is worth noting that the defined modified STOI function f is a derivative function, since each operation described above is differentiable. Therefore, we can optimize a modified STOI function f based loss by using backpropagation (BP) algorithm.

2.2. Proposed approach and loss function

Fig. 1 shows the diagram of the proposed approach. For the noisy speech enhancement, we employ the log magnitude spectrum of noisy speech as features to estimate the IRM, which is defined in equation (6) [4], and then apply the estimated ratio mask to the noisy magnitude spectrum to obtain the enhanced magnitude spectrum.

$$IRM(m,f) = \sqrt{\frac{X^2(m,f)}{X^2(m,f)+N^2(m,f)}} \tag{6}$$

where $X^2(m,f)$ and $N^2(m,f)$ denote the energy of clean speech and noise, respectively, at time frame m and frequency channel f.

To incorporate the temporal information, we utilize a context window to encompass the features from 2 frames before and 2 frames after the current frame. The ratio mask of the current frame is estimated by using this 5-frame context information. It should be pointed out that we only utilize one DNN to perform enhancement for each frame. Furthermore, there are many candidates for the denoising module. The only requirement is that the enhanced magnitude can be obtained by using the denoising module, since it is required by the modified STOI function computation.

After denosing, we can obtain the 24-frame enhanced magnitude spectrum $\boldsymbol{Y}^{24}$. Together with the corresponding 24-frame clean magnitude spectrum $\boldsymbol{X}^{24}$, we can compute the modified STOI value. Finally, at time frame m, the loss function is designed as follows,

$$\mathcal{L}(m) = (1 - f(\boldsymbol{X}_m^{24}, \boldsymbol{Y}_m^{24}))^2 + \lambda * \|\boldsymbol{X}_m^{24} - \boldsymbol{Y}_m^{24}\|_F/24 \tag{7}$$

where function f is the previous defined STOI function; $\|\cdot\|_F$ denotes the Frobenius norm; λ denotes a tunable hyper-parameter used to balance the two terms in the loss function. In our experiments, λ is set to 0.01. During training, we utilize a pre-trained ratio mask estimation neural network to initialize the denoising module in the proposed approach, and then train it by minimizing the proposed loss. During testing, the enhanced speech is synthesized by using the enhanced magnitude with the noisy phase. It should be pointed out that using the modified STOI function alone to design the loss function is not suitable, especially for the wide-band speech signal enhancement, because the STOI is only based information below the 4.3 kHz band. Consequently, we need to combine it with a MSE-based loss

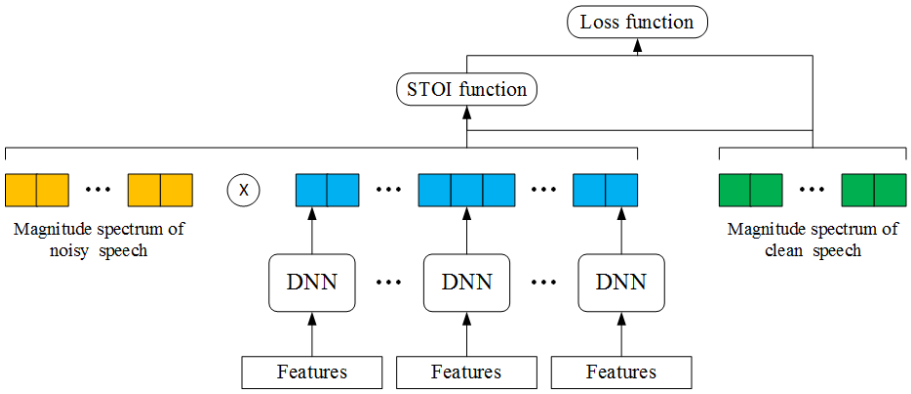

Fig. 1: Diagram of the proposed algorithm. Yellow rectangles denote the 24-frame magnitude spectrum of noisy speech; blue rectangles denote the 24-frame estimated ratio masks; green rectangles denote the corresponding 24-frame magnitude spectrum of clean reference speech. The 24-frame enhanced magnitude spectrum is obtained by applying the estimated ratio mask to the noisy magnitude spectrum. The DNNs that are used to do speech enhancement for each frame share the same parameters.

function in order to account for the whole speech spectrum.

Moreover, the computation of STOI values is based on 384 ms (24 frames in this study) temporal information. Therefore, optimizing the loss function (7) also explores the temporal context information at the output end. We note that such type of information is ignoring in multi-frame to one-frame supervised speech enhancement approaches, where the temporal information only at the input end is utilized by explicitly using a context window. Previous study [4] has shown that predicting neighbouring frames' target can bring us consistent improvements over predicting single frame target. Consequently, by using the output context information, the proposed loss function can potentially benefit for performing better enhancement.

3. EXPERIMENTAL SETUP

The proposed approach is evaluated using the IEEE corpus spoken by a female speaker [15], which consists 72 lists with 10 sentences in each list. List 1-50, List 67-72 and List 51-60 are used to construct training data, validation data and test data, respectively. Speech-shaped noise (SSN) and three types of non-stationary noise from NOISEX database [16] including speech babble (Babble), factory floor noise (Factory) and destroyer engine room noise (Engine) are used to generate noisy speech in our study. Each noise segment is 4 min long. The first 3 min is used for training and validation and the remaining is used for testing. For training/validation set, each clean sentence is mixed with 10 random noise segments at three SNR levels, namely, -5, 0 and 5 dB; for test set, each clean sentence is mixed with 1 random noise segment at five SNR levels, namely, -5, -3, 0, 3 and 5 dB, where -3 and 3 dB SNR conditions are unseen in the training set. Therefore, there are 500×4 (noise types)×3 (SNRs)×10 (noise segments)=60 k utterances in the training set; 50×4 (noise types)×3 (SNRs)×10 (noise segments)=6 k utterances in the validation set; 100×4 (noise types)×5 (SNRs)×1 (noise segment)=2 k utterances in the test set. Neither the sentences nor the noises in the test set are seen during training.

The proposed approach is first compared with a DNN-based masking denoising approach (**masking**), which employs a DNN to predict the IRM and utilizes the estimated ratio mask to perform denoising. It is also used as the denoising module in the proposed approach. Since part of the designed loss function is similar to that defined in the signal approximation approach (**SA**), we also compare our approach with the SA approach. The pre-trained masking model is utilized to initialize the SA model. To show that the proposed approach can be considered as a framework to improve the existing supervised speech enhancement approaches, we simply replace the denoising module with a DNN-based mapping approach (**mapping**), which is trained to learn a mapping function from log magnitude spectrum of noisy speech to that of clean speech. We denote this approach as "**mapping+proposed loss**". The normal mapping denoising approach is considered as a baseline to compare.

All DNNs in our study have three hidden layers with 1024 exponential linear units (ELUs) [17] in each layer. They are trained by using Adam [18] optimizer with dropout regularization [19]. The dropout rate in the experiments is set to 0.3. We employ sigmoid activation units in the output layer for the ratio mask estimation whose value is bounded between 0 and 1; otherwise, linear activation units are used. The input features are normalized to zero mean and unit variance. For the mapping approach, the training target is also normalized by using mean and variance normalization as suggested in [3]. The enhanced time-domain signal is synthesized by using noisy phase.

4. EVALUATION RESULTS

In our study, STOI, PESQ and SDR [20] are used to evaluate speech intelligibility and sound quality. Table 1 and Table 2 show the average performance of these three metrics under four types of noise with matched SNR levels (-5, 0 and 5 dB) and mismatched SNR levels (-3 and 3 dB), respectively. Boldface numbers highlight the best result under each condition.

Compared with the unprocessed noisy speech condition, each supervised speech enhancement approach improves the STOI, PESQ and SDR performance significantly, in both matched and mismatched SNR conditions. In other words, all the approaches investigated in this study can generalize well to the SNR conditions that are not included in the training data.

Since the objective of our study is to improve speech intelligibility, we focus on comparing the STOI scores of the different speech enhancement approaches first. As expected, the proposed approach achieves the best STOI score for each noise type. The performance trends of different approaches are similar under the four types of noise. Taking the Babble noise for an example, our approach out-

	STOI (in %)				**PESQ**				**SDR** (dB)			
	SSN	Babble	Factory	Engine	SSN	Babble	Factory	Engine	SSN	Babble	Factory	Engine
unprocessed	71.23	68.63	67.96	73.11	1.179	1.284	1.086	1.243	0.18	0.14	0.12	0.14
mapping	81.00	77.42	79.44	86.29	2.077	1.876	2.099	2.391	5.74	5.42	7.10	8.76
masking	84.00	80.01	82.34	89.18	2.135	1.896	2.063	2.465	6.70	6.18	8.52	10.61
SA	84.70	80.91	82.90	89.08	**2.233**	1.988	**2.192**	**2.557**	**7.43**	**6.98**	9.12	11.29
mapping+proposed loss	83.22	79.70	81.65	88.13	2.041	1.868	2.024	2.361	5.60	5.28	7.16	8.96
proposed approach	**85.70**	**81.99**	**84.31**	**90.27**	2.202	**1.996**	2.136	2.525	7.36	6.87	**9.14**	**11.32**

Table 1: Average performance scores for different enhancement approaches. Results averaged over mixtures under matched SNR levels (-5 dB, 0dB and 5 dB).

	STOI (in %)				**PESQ**				**SDR** (dB)			
	SSN	Babble	Factory	Engine	SSN	Babble	Factory	Engine	SSN	Babble	Factory	Engine
unprocessed	71.33	68.69	68.09	73.30	1.172	1.271	1.084	1.247	0.14	0.13	0.12	0.12
mapping	82.02	78.40	80.40	86.77	2.104	1.886	2.117	2.394	5.88	5.59	7.33	8.84
masking	84.80	80.70	83.13	86.69	2.154	1.884	2.071	2.474	6.72	6.23	8.59	10.60
SA	85.48	81.56	83.59	89.46	**2.256**	**1.999**	**2.208**	**2.574**	**7.52**	**7.06**	9.25	11.32
mapping+proposed loss	84.05	80.61	82.42	88.53	2.069	1.871	2.033	2.357	5.78	5.40	7.39	9.04
proposed approach	**86.48**	**82.68**	**84.98**	**90.61**	2.229	**1.999**	2.149	2.542	7.46	6.94	**9.27**	**11.35**

Table 2: Average performance scores for different enhancement approaches. Results averaged over mixtures under mismatched SNR levels (-3 dB and 3 dB).

performs the masking approach by about 2%. In fact, more STOI improvements are observed at lower SNR levels, where speech intelligibility improvements become more important since the communications are challenging in very noisy environments. At -5 dB, compared with the masking approach, 3.01% STOI score improvements are obtained for Babble by our approach. The SA approach performs better than the masking approach but worse than the proposed approach. We should point out that the masking approach and the SA approach are already very strong benchmarks to compare with and represent the state-of-the-art supervised denoising approaches.

Moreover, after replacing the masking denosing module in our approach with the mapping approach, about 2% STOI score improvements over the normal mapping approach are obtained under each type of noise on average. This demonstrates the potential benefit of migrating many other supervised speech enhancement approaches to the perceptually guided framework. Further speech intelligibility improvements are expected.

It is worth noting that the improvements in speech intelligibility provided by the proposed approach are not coming at the expense of a degradation in speech quality. Based on PESQ and SDR, our approach shows comparable performance to the SA approach, and outperforms the masking approach. In our experiments, we find that the tunable hyper-parameter λ affects speech intelligibility and quality of the enhanced speech. Currently, we are using a fixed value during system training and the value is determined empirically. However, some preliminary experiments by using a simple automatically adaptive λ show that better speech intelligibility and quality can be obtained under some noisy conditions. Designing a strategy to tune the parameter λ automatically is one direction to explore for the future study.

5. CONCLUSION

In this paper, we have proposed a perceptually guided speech enhancement approach aiming to suppress noise and improve speech intelligibility. Different from the existing supervised speech enhancement approaches, we incorporate a speech intelligibility metric into the loss function. Systematic evaluation shows that the proposed approach improves speech intelligibility over the existing supervised speech enhancement approaches in a wide range of noisy conditions. Future research will focus on incorporating additional perceptual information into both the loss function and the enhancement approach in general to further improve the performance.

6. ACKNOWLEDGEMENT

The authors would like to recognize and acknowledge João Felipe Santos for his early explorations on the perceptually guided speech enhancement project during his summer internship at Starkey in 2016.

7. REFERENCES

[1] J. Li, L. Deng, Y. Gong, and R. Haeb-Umbach, "An overview of noise-robust automatic speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, pp. 745–777, 2014.

[2] J. Ming, T. J. Hazen, J. R. Glass, and D. A. Reynolds, "Robust speaker recognition in noisy conditions," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, pp. 1711–1723, 2007.

[3] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, pp. 7–19, 2015.

[4] Y. Wang, A. Narayanan, and D. L. Wang, "On training targets for supervised speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, pp. 1849–1858, 2014.

[5] F. Weninger, J. R. Hershey, J. Le Roux, and B. Schuller, "Discriminatively trained recurrent neural networks for single-channel speech separation," in *IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, 2014, pp. 577–581.

[6] H. Erdogan, J. R. Hershey, S. Watanabe, and J. Le Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 708–712.

[7] Y. Wang and D. L. Wang, "A deep neural network for time-domain signal reconstruction," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 4390–4394.

[8] J. Le Roux, E. Vincent, and H. Erdogan, "Learning based approaches to speech enhancement and separation," in *INTERSPEECH Tutorials*, 2016.

[9] Y. Zhao, Z.-Q. Wang, and D. L. Wang, "A two-stage algorithm for noisy and reverberant speech enhancement," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 5580–5584.

[10] D. S. Williamson, Y. Wang, and D. L. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 24, pp. 483–492, 2016.

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

[12] 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 *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 81–85.

[13] 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 *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2001, vol. 2, pp. 749–752.

[14] V. Emiya, E. Vincent, N. Harlander, and V. Hohmann, "Subjective and objective quality assessment of audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, pp. 2046–2057, 2011.

[15] E. H. Rothauser, W. D. Chapman, N. Guttman, K. S. Nordby, H. R. Silbiger, G. E. Urbanek, and M. Weinstock, "IEEE recommended practice for speech quality measurements," *IEEE Transactions on Audio Electroacoust*, vol. 17, pp. 225–246, 1969.

[16] A. Varga and H. J. M. 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," *Speech communication*, vol. 12, pp. 247–251, 1993.

[17] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (ELUs)," *arXiv preprint arXiv:1511.07289*, 2015.

[18] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[19] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting.," *Journal of Machine Learning Research*, vol. 15, pp. 1929–1958, 2014.

[20] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE transactions on audio, speech, and language processing*, vol. 14, pp. 1462–1469, 2006.