A Priori SNR Estimation Based on a Recurrent Neural Network for Robust Speech Enhancement

Yangyang Xia[1], *Richard M. Stern*[1,2]

[1]Department of Electrical and Computer Engineering, Carnegie Mellon University
[2]Language Technologies Institute, Carnegie Mellon University

yangyanx@andrew.cmu.edu, rms@cs.cmu.edu

Abstract

Speech enhancement under highly non-stationary noise conditions remains a challenging problem. Classical methods typically attempt to identify a frequency-domain optimal gain function that suppresses noise in noisy speech. These algorithms typically produce artifacts such as "musical noise" that are detrimental to machine and human understanding, largely due to inaccurate estimation of noise power spectra. The optimal gain function is commonly referred to as the ideal ratio mask (IRM) in neural-network-based systems, and the goal becomes estimation of the IRM from the short-time Fourier transform amplitude of degraded speech. While these data-driven techniques are able to enhance speech quality with reduced artifacts, they are frequently not robust to types of noise that they had not been exposed to in the training process. In this paper, we propose a novel recurrent neural network (RNN) that bridges the gap between classical and neural-network-based methods. By reformulating the classical decision-directed approach, the *a priori* and *a posteriori* SNRs become latent variables in the RNN, from which the frequency-dependent estimated likelihood of speech presence is used to update recursively the latent variables. The proposed method provides substantial enhancement of speech quality and objective accuracy in machine interpretation of speech.

Index Terms: robust speech enhancment, *a priori* SNR estimation, decision-directed, recurrent neural networks.

1. Introduction

Speech enhancement (SE) has been one of the enabling technologies for robust speech processing applications for decades. SE algorithms strive to improve speech quality and intelligibility of speech signals degraded by additive noise [1]. Enhanced speech signals will benefit subsequent human listening experience or performance of machine tasks, such as automatic speech recognition and speaker verification. Classical signal processing methods for SE typically work in the frequency domain with optimization criteria associated with the spectral component of the enhanced speech. The technique ranges from heuristically estimating the power spectra [2], finding a linear filter that optimizes the mean squared error of the complex spectra [3], to minimum mean-squared error estimators (MMSE) that optimizes the (log) short-time spectral amplitude (STSA) [4, 5].

A priori signal-to-noise ratio (SNR) and *a posteriori* SNR arise as two important concepts from the derivation of the MMSE-STSA estimator [4]. The *a priori* SNR can be understood as the true instantaneous power ratio between each spectral component of clean speech and noise, while the *a posteriori* SNR can be viewed as the instantaneous power ratio between each spectral component of observed noisy speech and noise. Within this framework, the optimal gain function in the STSA domain for the well-known methods such as spectral subtraction [2], Wiener filter [3], maximum likelihood (ML) estimator [6], and MMSE estimation [4] can all be expressed in terms of *a priori* and *a posteriori* SNRs [7]. The enhancement problem thus becomes *a priori* SNR and *a posteriori* SNR estimation problem. For estimating the *a priori* SNR, a closed-form maximum-likelihood method and a recursive "decision-directed" method are proposed [4]. For estimating the *a posteriori* SNR, or equivalently the noise power, the minima-controlled recursive-averaging (MCRA) algorithm can be employed [7, 8, 9]. Despite the robustness of the decision-directed approach even in highly nonstationary noise environments, inexact heuristics in the estimation procedure often produce artifacts called musical noise, which is sometimes even more detrimental to machine tasks and human listening experience than noisy speech.

Independent of the classical approaches, researchers in the neural network (NN) community formulate the speech enhancement task to be a supervised learning problem. Recognizing the ideal ratio mask (IRM) in the STSA domain as a better training target than clean signal power or magnitude spectra [10], various neural network architectures have been explored to learn the IRM for SE. Some examples are feedforward deep neural networks [11, 12], deep denoising autoencoders [13], and recurrent neural networks (RNN) with long short-term memory [14]. Although these NN-based SE algorithms work well under noise conditions that appear in the training set, they typically suffer from degraded performance in unseen noise types as they attempt to learn a nonlinear mapping between noisy speech and the IRM.

A fusion system that combines the robustness and interpretability of the classical approach and the learning ability of the NN approach is clearly desirable. One previous study that attempts this fusion [15] proposes a NN version of spectral subtraction by having dedicated NNs for estimating noise alone, noise in noisy speech, and the enhanced speech. Although their NN structure is reminiscent of spectral subtraction, our experiments show that the latent variables do not learn the intended representation. Others [16, **?**] have attempted to improve *a priori* SNR estimation using NNs, but their systems are shallow combinations of multiple approaches at the input and output levels.

We propose a novel RNN that addresses these issues. We slightly modify the decision-directed approach to form a recurrent estimation of both the *a priori* SNR and *a posteriori* SNR, eliminating the need to estimate noise explicitly. This reformulation leads to a ratio-based representation for all variables, which have already proven to be superior training targets for neural network learning [10]. Among them, the *a priori* SNR, *a posteriori* SNR, and the speech-presence likelihood ratio are interpreted as latent recurrent cells of a recurrent neural network. This enables us to insert feedforward NNs to learn parameters that are normally heuristically determined using classical approaches. In addition, we introduce a learning objective function that jointly optimizes the MSE of STSA as well as the frame-level speech-presence detection accuracy.

2. The Signal-to-noise Ratio Recurrent Neural Network (SNRNN)

Our signal-to-noise ratio recurrent neural network (SNRNN) consists of a slightly modified version of the classical decision-directed *a priori* SNR estimation and a neural network component. Throughout the discussion, we assume additive noise in the short-time Fourier transform (STFT) domain:

$$X[m,k] = S[m,k] + N[m,k] \tag{1}$$

where $X[m,k]$, $S[m,k]$, and $N[m,k]$ denote the STFT at time frame m and frequency bin k of the observed noisy speech, clean speech, and noise, respectively. The end goal is to seek for the optimal gain function or IRM in the STSA domain, $G[m,k]$, such that the clean speech estimate $\hat{S}[m,k]$ can be obtained from the modified STSA and the phase from the noisy input:

$$\hat{S}[m,k] = G[m,k]|X[m,k]|e^{j\angle X[m,k]} \tag{2}$$

The *a priori* SNR $\xi[m,k]$ is defined by the ratio of the expected value of clean speech power to the expected value of the noise power:

$$\xi[m,k] = \frac{E[|S[m,k]|^2]}{E[|N[m,k]|^2]} \tag{3}$$

The *a posteriori* SNR $\gamma[m,k]$ is defined by the ratio of the instantaneous noisy speech power to the expected value of the noise power:

$$\gamma[m,k] = \frac{|X[m,k]|^2}{E[|N[m,k]|^2]} \tag{4}$$

In estimating the *a priori* and *a posteriori* SNRs, we replace the expected values by the corresponding instantaneous values:

$$\hat{\xi}[m,k] = \frac{|\hat{S}[m,k]|^2}{|\hat{N}[m,k]|^2}, \hat{\gamma}[m,k] = \frac{|X[m,k]|^2}{|\hat{N}[m,k]|^2} \tag{5}$$

Assuming that $S[m,k]$ and $N[m,k]$ are statistically independent zero-mean complex Gaussian random variables, Eq. 1 implies an additive relationship in the spectral power domain:

$$E[|X[m,k]|^2] = E[|S[m,k]|^2] + E[|N[m,k]|^2] \tag{6}$$

which leads to the definition of $\xi[m,k]$ in terms of $\gamma[m,k]$:

$$\xi[m,k] = E[\gamma[m,k]] - 1 \tag{7}$$

The decision-directed approach [4] calculates $\hat{\xi}[m,k]$ by linearly averaging the past and present estimates of *a priori* SNR:

$$\hat{\xi}[m,k] = a\hat{G}^2[m-1,k]\hat{\gamma}[m-1,k] + (1-a)max\{\hat{\gamma}[m,k]-1,0\} \tag{8}$$

where $0 < a < 1$ is the weighting coefficient, and $max\{\cdot\}$ is the element-wise maximum operator that prevents the current estimate from going below 0. The gain function $\hat{G}[m,k]$ is expressed in terms of $\xi[m,k]$ depending on the method to be used [7]. We use the Wiener estimate solution [3, 7], because the partial derivative of $\hat{G}[m,k]$ with respect to $\hat{\xi}[m,k]$ does not involve potential division by zero, which would result in gradient explosion during training:

$$\hat{G}[m,k] = \frac{\hat{\xi}[m,k]}{\hat{\xi}[m,k]+1} \tag{9}$$

Noise estimation is needed to calculate $\hat{\gamma}[m,k]$ by definition. Acknowledging the importance of the decision-directed approach, we adopt the MCRA algorithm [7, 8, 9] for noise power estimation. Specifically, the speech-absence (H_0^k) hypothesis and speech-presence (H_1^k) hypothesis are assumed for each frequency bin k of each frame m of the noisy signal:

$$\begin{aligned} H_0^k &: |\hat{N}[m,k]|^2 = b|\hat{N}[m-1,k]|^2 + (1-b)|X[m-1,k]|^2 \\ H_1^k &: |\hat{N}[m,k]|^2 = |\hat{N}[m-1,k]|^2 \end{aligned} \tag{10}$$

where $0 < b < 1$ is the weighting coefficient. In other words, the noise power in a specific frequency bin is recursively updated by a fraction of signal power from the previous frame only if it is classified as speech-absent. This decision is made by thresholding the likelihood ratio of speech-presence uncertainty:

$$\Lambda[m,k] \triangleq \frac{P(X[m,k]|H_1^k)}{P(X[m,k]|H_0^k)} \tag{11}$$

The previous assumption that the noise and speech DFT coefficients are independent, complex, and Gaussian leads to:

$$\Lambda[m,k] = \frac{1}{1+\hat{\xi}[m,k]} e^{\frac{\hat{\xi}[m,k]}{1+\hat{\xi}[m,k]}\hat{\gamma}[m,k]} \tag{12}$$

In our system, we replace the hard threshold used in Eq. 10 by a soft threshold to enable gradient backpropagation. We also rewrite $\hat{\gamma}[m,k]$ as a recursive function, eliminating the notion of noise estimation completely. Finally, we introduce the neural network component, along with the loss function.

2.1. Recurrent *A Priori* and *A Posteriori* SNR Estimation

The noise update rule in Eq. 10 can be interpreted as a recurrent nonlinear activation function. Specifically, let δ be a hard threshold of the log-likelihood ratio of speech-presence uncertainty above which the noisy frame is classified as speech-present. The update rule can then be rewritten as:

$$\frac{|\hat{N}[m,k]|^2}{|\hat{N}[m-1,k]|^2} = \beta(\Lambda[m-1,k]) + (1-\beta(\Lambda[m-1,k]))\hat{\gamma}[m-1,k] \tag{13}$$

where $\beta(\Lambda[m,k])$ is the scaled and shifted unit step function:

$$\beta(\Lambda[m,k]) = b + (1-b)u[log(\Lambda[m,k]) - \delta] \tag{14}$$

To enable gradient backpropagation in our RNN, we propose two nonlinearities, sigmoid and piecewise-linear, that have non-zero gradients around the decision boundary δ to replace the unit step function:

$$\begin{aligned} \beta_{sig}(\Lambda[m,k]) &= b + (1-b)\frac{1}{1+e^{-(log(\Lambda[m,k])-\delta)}} \\ \beta_{pwl}(\Lambda[m,k]) &= min\{1, max\{b, \tfrac{1-b}{2\epsilon}[log(\Lambda[m,k]) - (\delta-\epsilon)] + b\}\} \end{aligned} \tag{15}$$

where ϵ is a small positive constant that controls the width of the linear region. Combining the new update Eq. 13 with Eq. 5 we obtain:

$$\hat{\gamma}[m,k] = \frac{|X[m,k]|^2}{|X[m-1,k]|^2}\frac{\hat{\gamma}[m-1,k]}{\beta + (1-\beta)\hat{\gamma}[m-1,k]} \tag{16}$$

where β is shorthand for $\beta(\Lambda[m-1,k])$. Equations 8, 12, and 16 complete the recurrent estimation of both *a priori* and *a posteriori* SNR, without the need to explicitly estimate noise power. This distinction is important from the neural network learning perspective, as estimating a ratio mask rather than direct signal is desirable [10]. We now present the full system.

2.2. RNN for *A Priori* SNR Estimation

The recurrent structure described in the previous subsection naturally lends itself to a recurrent neural network framework. Specifically, we place a feedforward neural network immediately after each weighting factor, so that the RNN can learn the recursive averaging coefficients rather than applying heuristics:

$$\begin{aligned}\hat{\xi}[m,k] &= \hat{a1}_{m,k}\hat{G}^2[m-1,k]\hat{\gamma}[m-1,k] \\ &\quad + \hat{a2}_{m,k} max\{\hat{\gamma}[m,k]-1,0\} \\ \hat{\gamma}[m,k] &= \frac{|X[m,k]|^2}{|X[m-1,k]|^2}\frac{\hat{\gamma}[m-1,k]}{\hat{b1}_{m,k}+\hat{b2}_{m,k}\hat{\gamma}[m-1,k]} \\ \hat{a1}_{m,k} &= FF(a[m,k]), \hat{a2}_{m,k} = FF(1-a[m,k]) \\ \hat{b1}_{m,k} &= FF(\beta(\Lambda[m,k])), \hat{b2}_{m,k} = FF(1-\beta(\Lambda[m,k]))\end{aligned} \tag{17}$$

where $FF(\cdot)$ represents a feedforward neural network. Although the equations look very similar to Eqs. 8 and 16, we note two key differences. First, each coefficient is now parametrized by both time and frequency. Because of the interconnection of the neural network, these coefficients depend not only on the frequency bin they belong to, but also all other frequency bins. This is a useful generalization that is hard to carry out systematically in the classical framework due to the lack of a closed-form solution, but is easily realized in the neural network framework. Second, the heuristic constraint that weighting coefficients add up to 1 is removed.

The loss function of the SNRNN is twofold. We adopt the mean squared error in the STSA domain of the enhanced speech, not only because it is a popular objective function in deep learning SE methods, but also because it is the principle upon which the *a priori* SNR estimation is derived [4]:

$$E_{mse}[m] = \frac{1}{K}\sum_{k=0}^{K-1}(|S[m,k]| - |\hat{S}[m,k]|)^2 \tag{18}$$

In addition to the MSE-STSA loss, we introduce frame-level voice activity detection (VAD) loss. Because the recurrent structure is derived directly from the decision-directed approach, we can obtain the frame-level speech-presence log-likelihood ratio assuming statistical independence across frequency:

$$log\Lambda[m] = \sum_{k=0}^{K-1} log\Lambda[m,k] \tag{19}$$

where K is the total number of frequency bins. Assuming equal prior probability of speech presence and absence, the speech-presence probability given the noisy frame can be expressed as:

$$P(speech|X[m]) = \frac{\Lambda[m]}{1+\Lambda[m]} \tag{20}$$

We define the VAD loss for this two-class classfication problem as the cross entropy between the true and predicted speech-presence probability:

$$\begin{aligned}E_{vad}[m] &= -vad[m]log(P(speech|X[m])) \\ &\quad - (1-vad[m])log(1-P(speech|X[m]))\end{aligned} \tag{21}$$

where $vad[m] \in \{0,1\}$ is the true speech-presence probability for frame m. The overall objective function is:

$$E[m] = \alpha E_{mse}[m] + (1-\alpha)E_{vad}[m] \tag{22}$$

where $0 \leq \alpha \leq 1$ is the weighting coefficient.

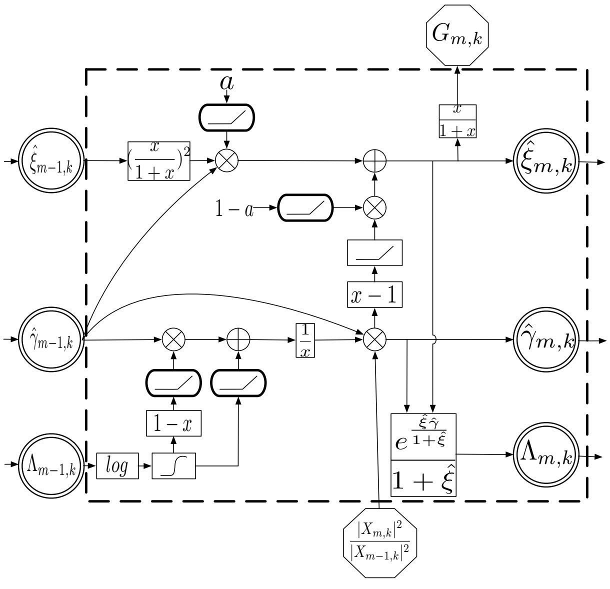

Figure 1: *SNRNN computation at frame m in the dashed box. Octagons hold input and output. Latent variables are bounded by rings. Feedforward networks are highlighted by bold capsules. Rectangles and circles are element-wise operations.*

We conclude our description of the SNRNN from the neural network perspective. From the zoomed-in view of the recurrent structure shown in Fig. 1, *a priori* SNR, *a posteriori* SNR, and speech-presence likelihood ratio are interpreted as latent variables that carry information across time frames. The four feedforward neural networks interact with instantaneous power ratios rather than direct speech or noise power, which potentially makes the system robust against unseen noise types. In fact, all variables in the network are represented as ratios, motivated by the finding that ratio masks are superior learning targets [10].

3. Experimental Results and Discussions

We conducted experiments using the RATS speech activity detection dataset [17] and the TIMIT dataset [18]. We selected the RATS dataset to train our system because it contains extremely challenging noise conditions. It also contains ample examples of both speech-present and speech-absent regions that are needed for training. To demonstrate the enhancement quality of our system, we choose speakers from four of the RATS channels for a speaker verification (SV) evaluation. To demonstrate the robustness of our system against other unseen noise types, we performed the global and local signal-to-distortion ratio (SDR) test [19] on the speech segments from the TIMIT dataset with digitally added noise samples taken from the NOIZEUS dataset [20]. In both experiments, we compared the SNRNN's performance (denoted NN in all tables and figures) with the classical Wiener solution using decision-directed *a priori* SNR estimation with MCRA noise estimation (denoted DD).

To train the SNRNN, we used a total 56 hours of 320 audio recordings sampled at 16 kHz from Channels A and H in the development partition of the RATS SAD dataset. For the SV task, we also used Channel D in training. During the training phase, 1000 320-ms audio segments were randomly sampled from all recordings to form one minibatch. We used oracle VAD information to maintain approximately equal numbers of speech-present and speech-absent frames within each minibatch. Next, we computed the STFT of each segment with a 32-ms Hamming

Table 1: *Improvement of SDR and 512-point Segmental SDR on 6300 TIMIT Utterances*

SNR (dB)	Cafeteria Babble				Train				Flight				Car			
	SDR		SegSDR		SDR		SegSDR		SDR		SegSDR		SDR		SegSDR	
	DD	NN	DD	NN	DD	NN	DD	NN	DD	NN	DD	NN	DD	NN	DD	NN
10	2.25	0.96	5.21	**8.20**	3.30	1.76	5.60	**8.30**	4.25	2.68	5.17	**8.53**	6.92	4.01	7.13	**10.5**
5	3.05	2.88	7.09	**9.43**	4.20	4.11	7.53	**9.76**	5.25	5.24	6.86	**9.90**	8.83	7.66	9.18	**12.7**
0	3.67	**3.90**	8.93	**10.4**	4.85	**5.61**	9.31	**11.0**	6.19	**6.88**	9.01	**11.4**	10.7	10.7	11.6	**15.3**
-5	3.44	**3.65**	9.54	**10.5**	5.27	**6.37**	10.7	**11.9**	7.00	**7.69**	11.3	**12.5**	12.3	**13.1**	13.8	**17.5**
-10	1.39	**1.66**	9.41	**9.80**	4.39	**5.57**	11.0	**12.3**	6.72	**7.24**	12.4	**13.1**	13.1	**14.5**	14.8	**18.4**
Mean	2.76	2.61	8.04	**9.67**	4.40	**4.68**	8.83	**10.7**	5.88	**5.95**	8.95	**11.1**	10.4	9.99	11.3	**14.9**

Table 2: *Speaker Verification Performance on RATS Speakers*

EER(%)	Noisy	DD	NN	Clean
Channel A	28.6	32.2	**24.9**	10.7
Channel B	36.6	37.2	**36.6**	11.5
Channel C	44.8	40.1	**36.7**	7.93
Channel H	43.2	29.7	**23.9**	10.8

window, 75% overlap between frames, and 512-point DFTs. Finally, we computed the magnitude STFT, retaining the first 257 frequency dimensions as the input to SNRNN.

For each neural network inside the SNRNN, we used a 3-layer feedforward network with 257 neurons in each of the input, hidden, and output layers. We used rectified linear units (ReLUs) as the activation function at all layers because SNRs are non-negative. We used the sigmoid function in Eq. 15. We chose constants $a = 0.98$, $b = 0.98$ and $\delta = 0.15$, which are effective for the classical DD approach [1]. The network parameters were initialized so that each network is an identity function prior to learning. $\hat{\gamma}$, $\hat{\xi}$, and Λ are initialized the same way as the decision-directed method [1]. $\alpha = 0.2$ is a good weighting constant for the loss function. We used stochastic gradient descent with a learning rate of 10^{-4} and a momentum of 0.9 to update all network weights.

We evaluated our enhancement system in terms of the equal error rate (EER) obtained for the SV task. The baseline SV system was trained using the ALIZE i-vector system setup described in [21]. Farsi speakers in the training partition of the RATS SAD dataset were used for evaluation. The enrollment consisted of 30, 28, 28, and 30 speakers from RATS Channels A, B, C, and H, respectively. 28, 35, 25, and 37 recordings from Channels A, B, C, and H, respectively, were tested against every enrolled speaker from their corresponding channels. Table 2 shows that the NN provides significant improvement for all channels except Channel B. The improvement for unseen channel C is even greater than the improvement for Channel A. In addition, NN provides better performance than DD in all cases. One notable finding in our experiment is that $\hat{G}[m, k]$ in Eq. 8 and 9 no longer need to be identical in SNRNN. The results in Table 2 were obtained using the power subtraction rule [7] for Eq. 8, and the Wiener filter rule for Eq. 9.

Table 1 shows the improvement of global and 512-point segmental SDR after applying DD and NN on noisy TIMIT utterances. The four types of noise we include are perceptually very different from the noise in RATS channels. Our results show that SNRNN consistently improves segmental SDR under all conditions, even though the global SDR sometimes is worse than DD in high SNRs (which are rare in our training data). We illustrate this phenomenon in Fig. 2, where we show compensated waveforms after DD and NN processing, respectively. The impulse-like speech waveform at around 3.8s is wrongly suppressed by NN. However, the burst of noise during 0-3s and 5-10s is better contained using NN. In addition, NN produces far fewer musical artifacts during 4-5s. The parameter α controls the tradeoff between speech smearing and noise suppression. Using MSE-STSA loss alone yields almost an identical system as the decision-directed approach, while using VAD loss alone results in a system that heavily suppresses noise and smears speech. Overall, the robustness of SNRNN processing was expected, as we note in Sec. 2, because the neural networks "see" only instantaneous SNRs.

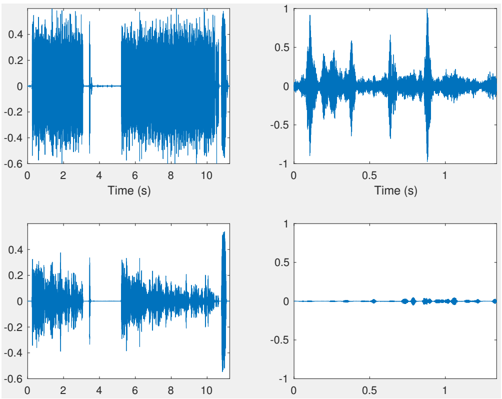

Figure 2: *Denoised waveforms. Top left and bottom left: denoised signal using DD and NN, respectively. Top right and bottom right: residual noise at 4-5s after 39.3dB amplification.*

4. Conclusions

In this paper, we have proposed a neural-network equivalent of the decision-directed *a priori* SNR estimation. We strongly advocate the use of instantaneous SNRs as internal representations in neural networks to accompany the use of IRMs as learning targets [10] for noise robustness. Our system preserves the robustness of the classical method while improving the accuracy of the recurrent approximations of *a priori* and *a posteriori* SNRs. Our results have shown that SNRNN processing can preserve speech and greatly suppress noise, while producing very few residual artifacts. In addition, our system can handle unseen nonstationary noise conditions when trained on very few noise types. We introduce the joint STSA-MSE and VAD loss function, and highlight the importance of VAD loss for balancing the level of noise suppression and speech distortion. In the future, we will attempt to improve the quality of enhanced speech in the speech-present regions, and extend the additive-noise framework to linear filtering for channel compensation.

5. Acknowledgments

We thank Benjamin Martinez Elizalde for useful suggestions on many prior drafts of this paper. We also thank our colleagues at the Afeka College of Engineering for their support.

6. References

[1] P. C. Loizou, *Speech Enhancement : Theory and Practice, Second Edition.* CRC Press, 2013.

[2] S. F. Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, no. 2, pp. 113–120, apr 1979.

[3] J. S. Lim and A. V. Oppenheim, "Enhancement and Bandwidth Compression of Noisy Speech," *Proceedings of the IEEE*, vol. 67, no. 12, pp. 1586–1604, 1979.

[4] Y. Ephraim and D. Malah, "Speech Enhancement Using a- Minimum Mean- Square Error Short-Time Spectral Amplitude Estimator," *IEEE Transactions on Acoustics Speech and Signal Processing*, vol. 32, no. 6, pp. 1109–1122, dec 1984.

[5] ——, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 33, no. 2, pp. 443–445, 1985.

[6] R. J. McAulay and M. L. Malpass, "Speech Enhancement Using a Soft-Decision Noise Suppression Filter," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 2, pp. 137–145, apr 1980.

[7] P. Scalart and J. Filho, "Speech enhancement based on a priori signal to noise estimation," in *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, vol. 2, 1996, pp. 629–632.

[8] J. Sohn and W. Sung, "A voice activity detector employing soft decision based noise spectrum adaptation," *Acoustics, Speech and Signal Processing, . Proceedings of the IEEE International Conference on.*, vol. 1, pp. 365–368, 1998.

[9] I. Cohen and B. Berdugo, "Noise Estimation by Minima Controlled Recursive Averaging for Robust Speech Enhancement," *IEEE SIGNAL PROCESSING LETTERS*, vol. 9, no. 1, 2002.

[10] 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, no. 12, pp. 1849–1858, 2014.

[11] Yong Xu, Jun Du, Li-Rong Dai, and Chin-Hui Lee, "A Regression Approach to Speech Enhancement Based on Deep Neural Networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 7–19, jan 2015.

[12] A. Kumar and D. Florencio, "Speech Enhancement In Multiple-Noise Conditions using Deep Neural Networks," 2016.

[13] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech Enhancement Based on Deep Denoising Autoencoder," *INTERSPEECH-2013*, pp. 436–440, 2013.

[14] F. Weninger, H. Erdogan, S. Watanabe, E. Vincent, J. Le Roux, J. R. Hershey, and B. Schuller, "Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9237. Springer, Cham, 2015, pp. 91–99.

[15] K. Osako, R. Singh, and B. Raj, "Complex recurrent neural networks for denoising speech signals," in *2015 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, WASPAA 2015*, 2015.

[16] S. Suhadi, C. Last, and T. Fingscheidt, "A data-driven approach to a priori SNR estimation," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 1, pp. 186–195, 2011.

[17] K. Walker, X. Ma, D. Graff, S. Stephanie, S. Stephanie, and K. Jones, "Rats speech activity detection ldc2015s02," Philadelphia: Linguistic Data Consortium, 2015.

[18] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1," *NASA STI/Recon Technical Report N*, vol. 93, Feb. 1993.

[19] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio Speech and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[20] Y. Hu and P. C. Loizou, "Subjective Comparison of Speech Enhancement Algorithms," *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, vol. 1, pp. I–I, 2006.

[21] A. Larcher, J.-F. Bonastre, B. Fauve, K. A. Lee, L. Christophe, H. Li, J. S. D. Mason, and J.-Y. Parfait, "ALIZE 3.0 - Open Source Toolkit for State-of-the-Art Speaker Recognition," *Interspeech*, no. August, pp. 2768–2772, 2013.