A STUDY OF TRAINING TARGETS FOR DEEP NEURAL NETWORK-BASED SPEECH ENHANCEMENT USING NOISE PREDICTION

Babafemi O. Odelowo, David V. Anderson

School of Electrical and Computer Engineering
Georgia Institute of Technology, Atlanta, GA
{bodelowo,anderson}@gatech.edu

ABSTRACT

Several architectures have been proposed for deep neural network (DNN)-based speech enhancement; however, these all utilize training targets related to the clean speech signal. In this paper, we evaluate the performance of several training targets in a noise-prediction DNN framework and compare the noise-prediction framework to a conventional speech-prediction network. Objective test results show that the mask-based targets are superior to the spectral magnitude target in the noise-prediction framework. The results also show that the best noise target outperforms the speech-prediction network in terms of objective quality and intelligibility metrics in seen noise conditions. The noise target is also competitive in unseen noise conditions, performing slightly worse in objective quality, but outperforming the speech-based target in objective intelligibility.

Index Terms— Speech enhancement, deep neural networks, noise estimation, speech quality, speech intelligibility

1. INTRODUCTION

Speech enhancement, the task of improving the quality and intelligibility of speech degraded by additive noise, has occupied the attention of the signal processing community for several decades due to its importance in applications such as personal and mobile communications, design of hearing aids, and robust automatic speech recognition (ASR) systems [1, 2].

While several algorithms including spectral subtraction [3], Wiener filtering [1,4], and minimum mean-square error (MMSE) algorithms [5–7] have previously been proposed for speech enhancement, there has recently been a focus on the use of data-driven methods. Techniques such as independent component analysis [8], non-negative matrix factorization (NMF) [9, 10], and deep neural networks (DNNs) [11, 12] have been used in speech enhancement frameworks. DNNs, in particular, have shown good generalization performance in several challenging acoustic conditions and can be considered to provide state-of-the-art performance [13].

DNN-based speech enhancement models are regression models that learn a mapping between noisy speech input features and a desired target. Common paradigms include spectral mapping [11, 14], time-frequency (T-F) masking [12, 13, 15], and multitask learning approaches [16, 17]. In the spectral mapping approach, the neural network predicts clean log power (or magnitude) spectra from noisy log spectra. T-F masking approaches, such as the ideal binary mask (IBM) and ideal ratio mask (IRM), use a neural network to estimate a T-F weighting function from noisy input features, and multitask learning approaches use a neural network to jointly estimate clean log power spectra and other secondary features such as mel-frequency cepstral coefficients (MFCCs), binary mask targets, and signal-to-noise ratio (SNR).

One notable feature of all the aforementioned approaches is that they utilize training targets based on clean speech features. Predicting clean speech features of low-SNR signals can be difficult as the speech may contain several noise-like, weak-energy segments in which the speech signal is dominated by the noise. The clean speech estimates in such segments are severely degraded as it is extremely challenging for a DNN to distinguish between speech and noise since the noisy speech is very similar to the pure noise, and the enhanced speech is consequently degraded. [16].

Recently, we proposed noise-prediction and time-domain subtraction framework as an alternate approach to DNN-based speech enhancement [18]. The rationale behind the use of the noise prediction approach was that learning a mapping between noisy speech input and added noise target features should be easier than learning a mapping between noisy speech input and the clean speech target features when the noise dominates the speech signal. The unexpected and somewhat contradictory performance of this approach, exhibiting stronger performance enhancing high-SNR signals than enhancing low-SNR signals, as well as the poor performance in unseen noise, indicated that the use of more robust features would be beneficial.

In this paper, we evaluate the performance of different training targets for DNN speech enhancement based on noise prediction. Three training targets are examined, and their performance is compared to that of a DNN trained with a conventional clean speech target. The spectral mapping framework commonly referred to as noise-aware training (NAT) [2, 11] is used for this comparison. The rest of the paper is organized as follows: the relation of this work to others in the literature is discussed in the next section. An overview of the proposed noise prediction systems is given in Section 3, experiments are described in Section 4, results are presented in Section 5, and conclusions are presented in Section 6.

2. RELATION TO PRIOR WORK

While there are several works investigating training targets for DNN-based speech enhancement [11–17], these have all been based on the prediction of clean speech and not noise targets. An in-depth study of training targets for supervised speech separation was conducted by Wang *et al.* [13]; however, this is the first study of this type for noise-estimation neural networks. Several factors that make the choice of training target for noise-estimation networks different than in speech-estimation networks are discussed. This work extends our initial work on the noise-prediction architecture by investigating more robust training targets and comparing the re-

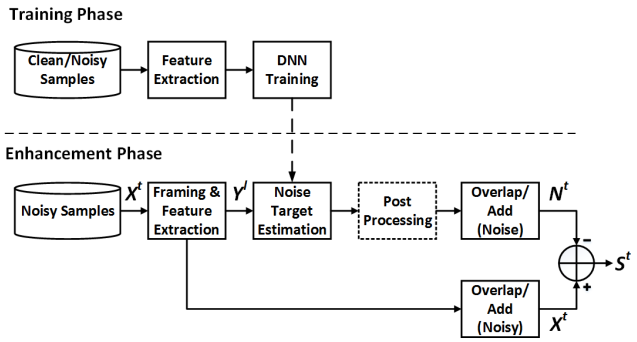

Figure 1: Block diagram of the proposed systems.

sults obtained with these targets to those obtained with conventional speech-estimation networks.

3. SYSTEM OVERVIEW

A block diagram of the proposed speech enhancement systems is shown in Figure 1. In the training phase, input-output feature pairs are extracted from the framed noisy speech and added noise signals respectively. Log magnitude spectral features are used as input features. Three training targets, namely, log spectral magnitude (LogFFT), Fourier magnitude spectrum mask (FFT-MASK), and a target which we introduce, the noise ratio mask (NRM), are evaluated.

1. Log Magnitude Spectrum

The magnitude of the short-time Fourier transform (STFT) spectrum of the noise is the natural choice for a training target in order to reconstruct the added noise. The STFT magnitude spectrum has a wide dynamic range, hence it is log compressed to reduce dynamic range and ease the DNN training process.

2. Fourier Magnitude Spectrum Mask

In conventional (speech prediction) spectral mapping models, the magnitude of the log spectral target is independent of the SNR of the noisy input signal since the target is the clean speech spectrum. The log spectral noise target, however, varies with SNR since the energy of the added noise depends on the SNR of the noisy input signal. The variation in the training target can be reduced by normalizing the magnitude spectrum of the added noise with that of the noisy speech signal. This gives the magnitude spectrum mask which is defined as:

$$M_{FFT}(t,\omega) = \frac{N(t,\omega)}{X(t,\omega)}, \tag{1}$$

where $M_{FFT}(t,\omega)$ is the mask, and $N(t,\omega)$ and $X(t,\omega)$ are the spectral magnitudes of the added-noise and noisy speech signals respectively. FFT-MASK is unbounded above, hence we enforced an upper bound to allow for more consistent training of the DNN. An upper bound of 3 was chosen by examining the distribution of a large random sample of the frequency bins.

3. Noise Ratio Mask

The noise ratio mask is defined as:

$$NRM(t,\omega) = \left(\frac{N^2(t,\omega)}{S^2(t,\omega)+N^2(t,\omega)}\right)^{\frac{1}{2}}, \tag{2}$$

where $N^2(t,\omega)$ and $S^2(t,\omega)$ represent the added-noise and speech signal power spectral densities respectively. The NRM is a bounded target with the range of [0,1], and can be seen to be equivalent to the frequency domain square-root Wiener filter if the speech and additive noise are assumed to be uncorrelated, and the noise is considered as the desired signal.

The network is trained by using the back-propagation algorithm to minimize a mean-square error criterion. Network parameters are updated using mini-batch stochastic gradient descent with momentum. The error criterion is

$$E = \frac{1}{N}\sum_{i=1}^{N} ||\hat{\mathbf{T}}_i(\mathbf{y}_i,\mathbf{\Theta}) - \mathbf{T}_i||^2 + \frac{\lambda}{2}||\mathbf{W}||_2^2 \tag{3}$$

where $\mathbf{y}_i$ is the input to the network, $\mathbf{\Theta} = \{\mathbf{W},\mathbf{b}\}$, represents the weights and biases in the network, $\hat{\mathbf{T}}_i(\mathbf{y}_i,\mathbf{\Theta})$ is the output of the network, $\mathbf{T}_i$ is the desired training target, λ is the regularization coefficient, and N is the mini-batch size.

In the enhancement phase, log spectral features extracted from noisy speech frames are fed into the trained network, and the network computes an estimate of the desired training target vector. For the log spectral target, LogFFT, a post-processing step follows [18, 19]. The magnitude spectrum estimates are used to compute a time-frequency (T-F) mask which is computed as:

$$H(t,\omega) = \min\left\{\left(\frac{\hat{N}^2(t,\omega)}{X^2(t,\omega)}\right)^{\frac{1}{2}}, 1\right\}, \tag{4}$$

where $\hat{N}^2(t,\omega)$ and $X^2(t,\omega)$ represent the estimated noise and noisy speech signal power spectral densities respectively. The mask, (4), is computed by normalizing the estimated added-noise signal power by the noisy signal power and enforcing an upper bound of unity. The mask thus represents a probability or confidence that a bin contains noise. The enforced upper bound also serves to prevent distortions that could be caused by estimation errors. The new post-processed noise spectral estimates are then obtained as:

$$\hat{N}_{pp}(t,\omega) = H(t,\omega)X(t,\omega), \tag{5}$$

where $X(t,\omega)$ is the noisy speech magnitude spectrum.

For the FFT-MASK and NRM targets, the noise spectral estimates are obtained by multiplying the predicted mask by the magnitude spectrum of the noisy speech as:

$$\hat{N}(t,\omega) = \hat{M}_{TF}(t,\omega)X(t,\omega) \tag{6}$$

where M_{TF} represents either the FFT-MASK or NRM targets. The predicted spectra are combined with the noisy phase, and a time-domain additive noise signal estimate is synthesized using the overlap-add method [20]. A real-time system is implemented by using a separate overlap-add buffer for the synthesis of the noisy signal frames. The noise-free speech signal estimates are then obtained by subtracting the added-noise signal estimate from the noisy speech signal as shown in Figure 1.

4. EXPERIMENTS

All experiments were performed using recorded sentences from the IEEE Corpus [21] included with the NOIZEUS database [1]. The corpus is comprised of 72 lists, each of which contains 10 sentences. Our noise source was a database of 100 non-speech sounds [22].

Noise	Description	Noise	Description
n1	Crowd	n6	Water
n2	Machine	n7	Wind
n3	Alarm/Siren	n8	Bell
n4	Traffic/Car	n9	Cough
n5	Animal	n10	Clap

Table 1: Description of noise types used in testing.

Both the noise-free speech and noise recordings were resampled to 8kHz. The training datasets were comprised of sentences taken from lists 1 - 60, while testing was done with the 50 sentences from lists 68 - 72.

Four training datasets were created by adding noise to the clean speech sentences. The first three datasets, which include those made for the NAT, FFT-MASK, and NRM models, were created by adding 50 noise types to the chosen clean speech samples at six SNR levels ranging from 20dB to -5dB in 5dB steps. The length of each dataset was about 50 hours. A similar-length training dataset was also created for the log magnitude spectral target. This dataset was, however, created by added the noise at seven SNR levels ranging from 20dB to -10dB in 5dB steps instead. The additional SNR level was added to increase the number of training samples that had a strong representation of the added-noise signals.

The speech signals were divided into 32ms frames and spectral features extracted from the clean speech, noisy speech, and from the added-noise signals were used to create input-output pairs for training the networks. Fourier analysis was performed using a Hamming window. The proposed noise prediction models used log magnitude spectral input features and targets as described in section 3, while the NAT model used log power spectral features following the common practice.

To allow the networks to take advantage of temporal information, each input vector included adjacent time frames. Consequently, each input vector was constructed as

$$\mathbf{y}_i = [\mathbf{x}_{i-l}, \ldots, \mathbf{x}_i, \ldots, \mathbf{x}_{i+l}]. \tag{7}$$

Five context frames, i.e. $l = 5$, for a total input length of eleven frames, were used in the training and evaluation of the enhancement systems.

The spectral input vectors for the NAT model were created by appending an estimate of the noise in each utterance to the noisy signal spectral input (7). The noise estimate, $\hat{\mathbf{n}}_i$, was fixed for each utterance and was obtained by averaging the first five frames of noisy speech log spectra as

$$\hat{\mathbf{n}}_i = \hat{\mathbf{n}} = \frac{1}{K} \sum_{k=1}^{K} \mathbf{x}_k. \tag{8}$$

The neural network models were all deep networks with three hidden layers, each containing 2000 hidden nodes. The hidden layers of all the networks used the rectified linear unit (ReLU) activation functions [23, 24], and the output layers were linear. Weights and biases of all the layers were initialized following the method of He *et al.* [25], and the networks were trained using gradient descent with momentum. The initial learning rate was set to 0.001 for the first 10 epochs, and then decreased by 10% every subsequent 10 epochs. The value of the regularization coefficient was set to 0.0001, and the the momentum coefficient was 0.9. A mini-batch size of 128 samples was used, and the networks were trained for 50 epochs. All networks were implemented and trained using the TensorFlow library [26].

Testing was done using both seen and unseen noise types. Ten noise types were used in each of the testing scenarios. In the seen noise tests, each of the noise types used during the enhancement or evaluation phase was one of the noise types used during the training phase. Conversely, in unseen noise testing, each of the noise types used during the evaluation phase had not been used during the training of the network. A description of the noise types is given in Table 1.

Speech quality and intelligibility were objectively evaluated using the perceptual evaluation of speech quality (PESQ) [27] and short-time objective intelligibility (STOI) [28] metrics respectively. PESQ scores range from -0.5 to 4.5 while STOI scores range from 0 to 1. These measures have been shown to have high correlation with subjective listening tests [29, 30] .

5. RESULTS

5.1. Evaluation in Seen Noise

The average PESQ scores for all the models in seen noise conditions are shown in Table 2. The LogFFT and FFT-MASK models are similar in performance, but FFT-MASK has a slight edge when average SNR is above 5dB, and LogFFT has a slight advantage otherwise. The overall average scores for both methods are basically equivalent. The NRM performs better than both LogFFT and FFT-MASK at all SNR levels and is the best of the noise prediction models in enhancing speech quality.

The average STOI scores for all the models in seen noise conditions are shown in Table 3. LogFFT performs better than FFT-MASK at all input SNR levels and has a average STOI score that is about 1.5% higher. The difference in performance between these two training targets also increases as SNR reduces. The LogFFT model also performs slightly better than the NRM model, however, with an average STOI difference that is always less than 1% , the difference can be seen to be insignificant. Considering both the PESQ and STOI scores, the NRM model performs best in seen noise conditions, followed by the LogFFT and then the FFT-MASK models.

SNR (dB)	Noisy	NAT	LogFFT	FFT-MASK	NRM
20	3.027	3.506	3.686	3.720	**3.765**
15	2.701	3.394	3.481	3.511	**3.590**
10	2.380	3.265	3.240	3.258	**3.380**
5	2.072	3.114	2.982	2.975	**3.134**
0	1.791	**2.932**	2.708	2.665	2.845
-5	1.503	**2.708**	2.409	2.327	2.513
AVG.	2.246	3.153	3.084	3.076	**3.205**

Table 2: Average PESQ scores for the different training targets and the noise aware training (NAT) models in seen noise conditions. The average over all SNR levels is denoted AVG.

5.2. Evaluation in Unseen Noise

The average PESQ scores for all the models in unseen noise conditions are presented in Table 4. Unlike in seen noise conditions,

SNR (dB)	Noisy	NAT	LogFFT	FFT-MASK	NRM
20	0.961	0.937	**0.981**	0.974	0.977
15	0.926	0.928	**0.968**	0.958	0.962
10	0.872	0.916	**0.947**	0.934	0.941
5	0.799	0.897	**0.917**	0.899	0.910
0	0.708	0.872	**0.874**	0.851	0.868
-5	0.608	**0.834**	0.817	0.787	0.808
AVG.	0.812	0.897	**0.917**	0.901	0.911

Table 3: Average STOI scores for the proposed and NAT systems in seen noise conditions.

FFT-MASK performs better than LogFFT at all SNR values and is consequently better on average. The NRM model is once again better than both the FFT-MASK and LogFFT models and is the best of the noise prediction models in enhancing speech quality.

The average STOI scores for all the models in unseen noise are shown in Table 5. FFT-MASK performs slightly better than LogFFT, but the average STOI difference is insignificant. The NRM model outperforms both the FFT-MASK and LogFFT models with a difference of about 2% in the average STOI scores.

Considering the performance of all the noise prediction models in both seen and noise noise conditions, the NRM model preforms best, followed by the FFT-MASK, and lastly, the LogFFT models. The two normalized models, NRM and FFT-MASK, perform markedly better than LogFFT in unseen noise conditions. This could be because their training targets are related to both the added-noise and the noisy signal spectra, and the noisy signal spectrum, in effect, constrains the value of the target. In unseen noise conditions, the constraining effect remains and the targets generalize better. This is not the case with LogFFT, and it is therefore more susceptible to prediction errors in unseen noise conditions.

SNR (dB)	Noisy	NAT	LogFFT	FFT-MASK	NRM
20	3.182	3.426	3.292	3.413	**3.530**
15	2.875	3.242	3.007	3.134	**3.266**
10	2.569	**3.020**	2.715	2.842	2.976
5	2.288	**2.760**	2.420	2.538	2.664
0	2.036	**2.475**	2.137	2.233	2.345
-5	1.779	**2.182**	1.865	1.942	2.032
AVG.	2.455	**2.851**	2.573	2.684	2.802

Table 4: Average PESQ scores for the proposed and NAT systems in unseen noise conditions.

SNR (dB)	Noisy	NAT	LogFFT	FFT-MASK	NRM
20	0.958	0.935	0.965	0.965	**0.970**
15	0.925	0.922	0.937	0.939	**0.949**
10	0.876	0.900	0.893	0.899	**0.915**
5	0.813	0.862	0.832	0.842	**0.863**
0	0.736	**0.804**	0.755	0.767	0.793
-5	0.650	**0.727**	0.666	0.677	0.706
AVG.	0.826	0.858	0.841	0.848	**0.866**

Table 5: Average STOI scores for the proposed and NAT systems in unseen noise conditions.

5.3. Comparison of Speech and Noise Prediction Models

The PESQ scores in Table 2 show that the noise-prediction models perform comparatively well at higher SNR values, and in seen noise conditions. The NRM model outperforms the NAT model at all SNR values above 0dB, however, the NAT model performs better at the lower SNR values. The difference between the PESQ scores of the NAT and noise models as SNR decreases is worse for the LogFFT and FFT-MASK models than it is for the NRM model. The likely reason for the observed drop in performance with SNR is that the training targets of the noise prediction models are SNR dependent. As such, the DNN might tend to average over these targets leading to under-estimation of the noise in low-SNR signals. Our informal listening tests confirmed that the low-SNR speech signals enhanced by the noise prediction models had more residual noise than those enhanced by the NAT model.

The STOI scores in Table 3 show that that the noise prediction models also perform well in enhancing intelligibility in seen noise conditions. The LogFFT model outperforms the NAT model at all SNR values except -5dB, and both the LogFFT and NRM models outperform the NAT model by about 2% on average.

The PESQ scores in Table 4 show that the NRM model performs slightly better than the NAT model above 10dB SNR in unseen noise conditions. 10dB SNR marks an inflection point at which the NAT model becomes increasingly better than the NRM model as SNR decreases, and the NAT model performs slightly better than the NRM model on average. The STOI scores in Table 5 show the NRM model performs better than the NAT model above 0dB SNR, but the performance margin reduces as SNR decreases. The average STOI score of the NRM model is about 1% better than that of the NAT model.

The noise models can thus be seen to perform comparatively well to the NAT model in enhancing speech quality at higher average SNR values and in enhancing intelligibility even when the latter is not accompanied by corresponding quality enhancement. The most likely reason for this observation lies in how target estimation errors differently affect both model types. Estimation errors in the NAT model could either attenuate or amplify portions of the speech signal spectrum and cause attending distortions in the enhanced speech. Amplification distortions of the enhanced speech spectrum have been shown to adversely affect the speech intelligibility [31]. Estimation errors could similarly affect the estimated noise spectrum, however, these are more likely to occur in noise-dominant speech segments and do not affect the enhanced speech spectrum. Our informal listening tests showed that while the lower-SNR signal enhanced by the NAT model tended to be garbled, this was not the case with the noise models.

6. CONCLUSION

A study of DNN training targets for noise prediction was conducted. Objective test results showed the noise models were particularly effective in enhancing the intelligibility of noisy speech signals. The mask-based noise targets, which inherently include a normalization factor, performed better than the spectral noise target in unseen noise conditions. The noise ratio mask was the best all-round noise target. It outperformed the NAT model in seen noise conditions and in improving intelligibility in unseen noise, but fell short at lower SNR values. In future work, we will investigate how to further improve the robustness of the noise models in low-SNR and in unseen noise conditions.

7. REFERENCES

[1] P. C. Loizou, *Speech enhancement: theory and practice*. CRC press, 2013.

[2] M. L. Seltzer, D. Yu, and Y. Wang, "An investigation of deep neural networks for noise robust speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 7398–7402.

[3] S. 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, 1979.

[4] 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.

[5] 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. 1109–1121, 1984.

[6] ——, "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.

[7] I. Cohen and B. Berdugo, "Speech enhancement for non-stationary noise environments," *Signal processing*, vol. 81, no. 11, pp. 2403–2418, 2001.

[8] L. Hong, J. Rosca, and R. Balan, "Independent component analysis based single channel speech enhancement," in *Signal Processing and Information Technology, 2003. ISSPIT 2003. Proceedings of the 3rd IEEE International Symposium on*. IEEE, 2003, pp. 522–525.

[9] K. W. Wilson, B. Raj, and P. Smaragdis, "Regularized non-negative matrix factorization with temporal dependencies for speech denoising," in *Ninth Annual Conference of the International Speech Communication Association*, 2008.

[10] N. Mohammadiha, P. Smaragdis, and A. Leijon, "Supervised and unsupervised speech enhancement using nonnegative matrix factorization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 2140–2151, 2013.

[11] 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, no. 1, pp. 7–19, 2015.

[12] Y. Zhao, D. Wang, I. Merks, and T. Zhang, "DNN-based enhancement of noisy and reverberant speech," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 6525–6529.

[13] Y. Wang, A. Narayanan, and D. 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.

[14] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal Processing Letters*, vol. 21, no. 1, pp. 65–68, 2014.

[15] 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 *International Conference on Latent Variable Analysis and Signal Separation*. Springer, 2015, pp. 91–99.

[16] T. Gao, J. Du, Y. Xu, C. Liu, L.-R. Dai, and C.-H. Lee, "Improving deep neural network based speech enhancement in low snr environments," in *International Conference on Latent Variable Analysis and Signal Separation*. Springer, 2015, pp. 75–82.

[17] S.-W. Fu, Y. Tsao, and X. Lu, "SNR-Aware convolutional neural network modeling for speech enhancement." in *INTERSPEECH*, 2016, pp. 3768–3772.

[18] B. O. Odelowo and D. V. Anderson, "A noise prediction and time-domain subtraction approach to deep neural network based speech enhancement," in *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, Dec 2017, pp. 372–377.

[19] ——, "A mask-based post processing approach for improving the quality and intelligibility of deep neural network enhanced speech," in *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, Dec 2017, pp. 1134–1138.

[20] L. R. Rabiner and R. W. Schafer, *Theory and applications of digital speech processing*. Pearson, 2011.

[21] E. Rothauser, W. Chapman, N. Guttman, K. Nordby, H. Silbiger, G. Urbanek, and M. Weinstock, "IEEE recommended practice for speech quality measurements," *IEEE Trans. Audio Electroacoust*, vol. 17, no. 3, pp. 225–246, 1969.

[22] G. Hu, "A corpus of nonspeech sounds," [Online; accessed 13-September-2016]. [Online]. Available: http://web.cse.ohio-state.edu/pnl/corpus/hunonspeech/hucorpus.html

[23] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks." in *Aistats*, vol. 15, no. 106, 2011, p. 275.

[24] M. D. Zeiler, M. Ranzato, R. Monga, M. Mao, K. Yang, Q. V. Le, P. Nguyen, A. Senior, V. Vanhoucke, J. Dean, *et al.*, "On rectified linear units for speech processing," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 3517–3521.

[25] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1026–1034.

[26] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015, software available from tensorflow.org. [Online]. Available: http://tensorflow.org/

[27] 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 *Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on*, vol. 2. IEEE, 2001, pp. 749–752.

[28] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*. IEEE, 2010, pp. 4214–4217.

[29] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Transactions on audio, speech, and language processing*, vol. 16, no. 1, pp. 229–238, 2008.

[30] 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, no. 7, pp. 2125–2136, 2011.

[31] P. C. Loizou and G. Kim, "Reasons why current speech-enhancement algorithms do not improve speech intelligibility and suggested solutions," *IEEE transactions on audio, speech, and language processing*, vol. 19, no. 1, pp. 47–56, 2011.