INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Perception Optimized Deep Denoising AutoEncoders for Speech Enhancement

Prashanth Gurunath Shivakumar, Panayiotis Georgiou

University of Southern California, Los Angeles, CA, USA

pgurunat@usc.edu, georgiou@sipi.usc.edu

Abstract

Speech Enhancement is a challenging and important area of research due to the many applications that depend on improved signal quality. It is a pre-processing step of speech processing systems and used for perceptually improving quality of speech for humans. With recent advances in Deep Neural Networks (DNN), deep Denoising Auto-Encoders have proved to be very successful for speech enhancement. In this paper, we propose a novel objective loss function, which takes into account the perceptual quality of speech. We use that to train Perceptually-Optimized Speech Denoising Auto-Encoders (POS-DAE). We demonstrate the effectiveness of POS-DAE in a speech enhancement task. Further we introduce a two level DNN architecture for denoising and enhancement. We show the effectiveness of the proposed methods for a high noise subset of the QUT-NOISE-TIMIT database under mismatched noise conditions. Experiments are conducted comparing the POS-DAE against the Mean Square Error loss function using speech distortion, noise reduction and Perceptual Evaluation of Speech Quality. We find that the proposed loss function and the new 2-stage architecture give significant improvements in perceptual speech quality measures and the improvements become more significant for higher noise conditions.

Index Terms: Speech enhancement, Perception optimized speech denoising auto-encoders, Deep Neural Networks, objective function, Denoising

1. Introduction

Speech signals encode a multitude of information and in addition to human consumption are used in a range of automated speech processing tasks such as automatic speech recognition, emotion recognition, voice activity detection, and speaker identification systems. Such systems require good quality input signals and thus speech enhancement systems that can reduce noise are often employed as a pre-processing step. Speech enhancement has also been used to increase speech intelligibility in adverse noise conditions.

Speech Enhancement algorithms can be broadly classified into 4 categories [1]: (i) *Spectral subtractive* algorithms work on the principle of estimating noise spectrum during non-speech regions and subtracting it from the speech [2] [3]; (ii) *Statistical-model-based* algorithms are based on the principle of stochastic estimation like Minimum Mean Square Error (MMSE) Estimation of Spectrum Amplitude [4], Maximum Likelihood Estimation of magnitude of Speech Spectrum [5] and Wiener filters [6]; (iii) *Subspace* algorithms assume the clean signals to be a subspace of the noisy signal. Linear algebra concepts like Singular Value Decomposition (SVD) [7] and Karhunen-Loeve transform (KLT) [8] are used to decompose the noisy speech signal to speech and noise; and, (iv) *Binary Masking* algorithms apply a binary mask to the time-frequency representation of noisy signal to eliminate certain frequency bins by applying a threshold on Signal-to-Noise Ratio (SNR) [9, 10]. Some spectral subtraction algorithms [2, 3] and Wiener filter based techniques [6] introduced artifacts often referred to as "musical noise". However MMSE based techniques were able to yield significantly reduced musical noise [4].

Recently, Neural Networks have shown significant gains in many fields including in speech processing. The ability of Neural Networks to model complex non-linear mapping functions make them suitable for denoising tasks, thus efforts are under way for applying neural networks for time-domain and transform-domain mappings [11]. Time domain mappings employ training neural networks directly to map the noisy speech to the clean speech. They assume that the hidden layer transformations allow for the separation of noise from the speech. The functional role of each of the layers has been studied in [12, 13]. The transform domain technique initially transforms the speech signal to a domain with more desirable perceptual or recognition properties. The neural network is then trained to map transformed noisy features to clean features which are then transformed back to speech. It has been shown that log-spectral features improve denoising [14]. Cepstral domain de-noising as a pre-processing module has also improved ASR performance [15]. Further, including additional parameters describing noise and speech along with the noisy signal have proven advantageous for denoising [14].

Neural Network architectures adopted for denoising are often referred to as *Denoising Auto-Encoders* (DAE). A DAE is an auto-encoder which attempts to map the noisy inputs to their clean versions. Architectures adopted for denoising have remained more or less the same, even though changes in training techniques like dropout [16, 17], noise aware training [17, 16, 18], greedy layer-wise pre-training with fine tuning [19] have proven to be beneficial. In *Noise Aware Training* (NAT), an estimate of noise is provided along with the input [16, 17, 18]. DAE has proved to be more robust and achieved significant improvements both in terms of subjective and objective evaluations without the presence of musical noise artifacts observed in typical speech enhancement denoising algorithms [20].

Recurrent Neural Network (RNN) denoising referred as Recurrent Denoising Autoencoder (RDAE) has also shown significant performance gains [21, 22, 23]. RDAE can exploit the temporal context information embedded in signals [21]. This is particularly advantageous for non-stationary noise conditions where traditional DAE performs poorly due to its inability to exploit temporal information. *Long Short-Term Memory* DNNs have also proven to perform better than RNN [21] for denoising.

In spite of the improvements provided by the DAE in application to speech enhancement, it is known to cause additional speech distortion due to over-smoothing and clipping of clean speech due to the global MSE objective function. This effects and limits the perceptual quality of denoised speech because

http://dx.doi.org/10.21437/Interspeech.2016-1284

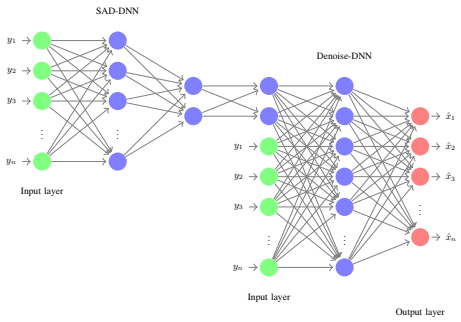

Figure 1: Type-I architecture

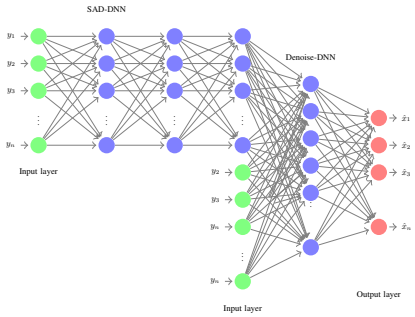

Figure 2: Type-II architecture

of the observed muffling effect. Post-processing techniques like global variance equalization are used to reduce the artifacts [16, 24, 25, 26].

In this paper, we propose *(i)* a new objective function to reduce over-smoothing problem and *(ii)* introduce a 2-stage DNN architecture to exploit speech activity information for speech enhancement. The rest of the paper structure is as follows. Sec. 2 describes in detail the proposed loss function and the introduced architecture changes. Our system details, database and evaluation criteria employed is provided in Sec. 3. Experimental results and analysis is presented in Sec. 4. Relation to prior work is discussed in Sec. 5 before concluding in Sec. 6.

2. Proposed System

2.1. Perception-Optimized Loss Function

DAE based speech enhancement techniques have shown immense improvements in the field of speech enhancement and are presently state-of-the-art. A traditional DAE uses a typical MSE loss function as an objective to find the mapping function between noisy and its clean version of the speech. The MSE objective error function to be minimized by the stochastic gradient algorithm is as follows:

$$E = \frac{1}{2}\|X - \hat{X}\|_2^2 \tag{1}$$

where E is the error, $\hat{X}$ is the output of the DNN during forward propagation, X is the target label (clean speech - log filter bank). The gradient of the error function to be back-propagated (assuming the output layer is linear) can be derived:

$$\frac{\partial E}{\partial \hat{X}} = \frac{\partial}{\partial \hat{X}}\frac{1}{2}(X - \hat{X})^2 = \hat{X} - X \tag{2}$$

The gradient of MSE is thus a linear function. The linear property of the gradient leads to the over-smoothing issue prevalent in traditional DAE. This is because, the penalty for clipping off a speech segment is same as the penalty for clipping off noise as far as the euclidean distance from the target clean speech is the same. This effect is more prevalent especially at lower SNRs i.e., when the noise is significantly high compared to the speech signals, the DAE clips off the speech segments at high noise regions to optimize its global error function.

In terms of perceptual quality it is better to preserve speech segments with residual error rather than clip speech segments to remove noise. Based on this we design a loss function which assigns high penalty against signal removal and retains the same MSE error for noise removal. Thus we propose a new, perceptually motivated, loss function:

$$E = \begin{cases} \frac{1}{2}\|X - \hat{X}\|_2^2 & \text{if } \hat{X} \geq X \\ \frac{1}{2}\|X - \hat{X} + p\|_2^2 & \text{if } \hat{X} < X \end{cases} \tag{3}$$

where p is the penalty assigned for speech clipping which is a positive scalar constant. The resulting error gradient (assuming linear output layer) is given by:

$$\frac{\partial E}{\partial \hat{X}} = \begin{cases} \hat{X} - X & \text{if } \hat{X} \geq X \\ \hat{X} - X - p & \text{if } \hat{X} < X \end{cases} \tag{4}$$

When the DNN output is greater than the clean speech signal (no clipping occurs), we use the traditional MSE objective and when the DNN output is less than the target clean speech (clipping occurs) we penalize by increasing the error by p. The DNN training with the proposed loss function learns to approximate the clean signal meanwhile avoids removing the desired signal. Note that the proposed loss function equals the MSE loss function when penalty is set to zero ($p = 0$).

2.2. Proposed DNN architecture

Prior studies have shown that addition of noise statistics as shown with *Noise Aware Training* (NAT) have yielded performance benefits [16, 17, 18]. However, correct estimation of the noise statistics used in NAT is critical.

Type I: Motivated by NAT, we propose a new 2 stage DNN architecture for denoising and speech enhancement for higher noise environments. The architecture involves a denoising stage preceded by *Speech Activity Detection* DNN (SAD-DNN). In this proposed framework, by providing the speech and non-speech region markings, we enable the neural network to implicitly compute the underlying noise statistics over the non-speech regions, and to exploit these towards improved denoising.

Type I-Module 1:The first stage comprises of a DNN trained to predict speech activity on each frame of audio. This SAD-DNN is trained using the cross-entropy loss function with softmax output layer as a classification problem to predict 0 for regions without speech and 1 for regions with speech.

Type I-Module 2:The second stage comprises of a DNN trained to map the noisy speech signals to their clean version. This is similar to the traditional DAE, with the following two exceptions (i) the output of the SAD-DNN becomes an additional input to this denoising DNN together with the noisy signal, (ii) the denoising DNN is trained using the newly proposed objective function (opposed to MSE).

Type I Disjoint and Joint Training: During training, the two stage system is trained as 2 independent DNN's optimized for their respective tasks. For the 2nd stage, the input features are augmented with noisy versions of the true VAD labels, v_1 and $v_2 = 1 - v_1$, approximating the soft-max outputs of the first stage system. A real system would never be certain, while our reference labels are binary 0s (non speech regions) or 1s (for

Train:	CAFE-FOODCOURTB-1, CAFE-FOODCOURTB-2, CAR-WINDOWNB-1, CAR-WINDOWNB-2, HOME-KITCHEN-1, HOME-KITCHEN-2, REVERB-POOL-1, REVERB-POOL-2, STREET-CITY-1, STREET-CITY-2.
Test:	CAFE-CAFE-1, CAFE-CAFE-2, CAR-WINUPB-1, CAR-WINUPB-2, HOME-LIVINGB-1, HOME-LIVINGB-2, REVERB-CARPARK-1, REVERB-CARPARK-2, STREET-KG-1, STREET-KG-2.

Group	SNR			Total
	Clean	-5dB	-10dB	
Train	1000	733	767	2500
Test	500	243	257	1000

Table 1: Database Summary: At the top we see the conditions used for training and testing while below we list the number of utterances employed as train and test sets.

speech regions). As a better initialization point we use the logarithm of uniformly distributed random values in the intervals [0,0.5] and (0.5,1] to indicate non-speech and speech regions respectively. Next, the two modules are concatenated and retrained jointly. Additionally, we replace the softmax output layer with a sigmoid layer to enable back-propagation. The concatenated DNN is fine-tuned by joint optimization of the two modules with the newly proposed loss function until convergence. We refer to this system, as in Fig. 1, as Type-I.

Type II: We also experiment on relaxing the information bottleneck between the two layers. The proposed Type I architecture has a bottleneck from module 1 to module 2, limiting effective back-propagation of the errors. To tackle this we propose a bottleneck expansion in the connectivity of the first and second stages. This proposed, Type II, architecture uses replication of the output layer of module 1 to expand the information flow to module 2 by widening the number of connections (to 50 dim.) as shown in figure 2, i.e., the 2-dimensional softmax layer from the module 1 is first replaced by sigmoids, and additional sigmoid units are concatenated to the two sigmoids initialized with identical weights and biases. The bottleneck between the 2 modules no longer exists. This enables for better back-propagation of errors and adds more flexibility during joint optimization.

3. Experimental Setup

3.1. Database

We test our system using the high noise subset of the QUT-NOISE-TIMIT corpus [27]. The QUT-NOISE-TIMIT corpus consists of 600 hours of noisy speech sequences created by mixing 10 hours of noise with TIMIT clean speech [28]. The noise database consists of 5 different background noise scenarios recorded at 10 unique locations. The resulting mixed speech sequences are classified into three categories: (i) low noise (15dB and 10dB SNRs), (ii) medium noise (5dB and 0dB SNRs) and (iii) high noise (-5dB and -10dB SNRs). In this study, only the high noise subset is considered. The noise scenarios are cafe, home, street, car and reverberation, each with 2 unique sessions and 2 unique locations. The database was divided into 2 parts for training and testing. The summary of the database and the division we used for our experiments are in Table 1. The noise environments used for training are as in Table 1 top, while the division of dataset in terms of number of utterances and noise levels are shown at the bottom. As can be seen the train and test are mutually exclusive.

3.2. System Description

Baseline: A typical DAE based speech enhancement system as used in [19, 20] forms our baseline. In our setup we use log-power spectral features due to its perceptual relevant properties [14] similar to the setup of [16, 18]. Log-power spectral features of dimension 257 was used along with left and right context of 4. Two hidden layers with sigmoid activation functions each of dimension 2000 were used for training. The number of hidden layers were restricted to 2 since having more layers was shown to provide insignificant improvements [19]. A linear output layer with MSE loss function was used.

POS-DAE: For module I, speech activity detection, we use a DNN with 2 hidden layers. The first hidden layer comprises of a Long Short-Term Memory (LSTM) with tanh activation functions and the second layer a feed-forward DNN with rectified linear units (RELU) both of dimension 50. Module II shares a similar architecture to the baseline system.

Implementation: We modified and used the customized version of the KALDI toolkit [29] for all our experiments. Details of the neural network training and algorithms used in KALDI can be found in [30]. The proposed custom objective function was implemented in CUDA to enable GPU implementation.

3.3. Speech Reconstruction

For evaluation purposes, to compute the perceptual evaluation of speech quality (PESQ), we need the waveform to be reconstructed from the log-power spectrogram features. To reconstruct, we apply the exponential operation to the log-power spectral features and take an inverse transformation using the phase signal derived from the noisy speech similar to [16, 19]. Using the phase information from the noisy speech for the reconstruction of the enhanced speech is justified because of the insensitivity of the human ear to small phase distortions in speech [11]. Finally the waveform is reconstructed using overlap-add synthesis. To make a fair assessment of the performance, the reference clean signal is decomposed and reconstructed in the same way.

3.4. Evaluation Measures

We present our results in terms of the objective measures of (i) *Noise Reduction*, (ii) *Speech Distortion* and (iii) *Perceptual Evaluation of Speech Quality* (PESQ) which are all well adopted in the speech enhancement community. Noise reduction (NR) and Speech Distortion (SD) are given by:

$$\text{NR} = \frac{1}{N \times d} \sum_{i=1}^{N} |\hat{X}_i - Y_i| \quad (5) \qquad \text{SD} = \frac{1}{N \times d} \sum_{i=1}^{N} |\hat{X}_i - X_i| \quad (6)$$

where N is the number of testing samples, d is the feature dimension, $\hat{X}_i$ is the enhanced output from DNN, X_i is the target clean speech features, Y_i is the noisy input features. PESQ is also adopted as it is said to have high correlation with the mean opinion score (MOS) [31]. PESQ is presented as a score between -0.5 to 4.5, where -0.5 and 4.5 represents lower bound and higher bound for speech quality respectively. Details of PESQ computation can be found in [31].

4. Experimental Results and Discussions

First, we present comparison of the effectiveness of the introduced loss function with identical architectures. Next, we motivate the proposed 2-stage architecture with an oracle Speech Activity Detection first layer. Then, we present results of the

Penalty		2	4	6	8	10
PESQ	-5dB	1.848	1.979	2.054	2.102	2.147
	-10dB	1.749	1.885	1.956	1.997	2.025

Table 2: Effect of penalty on PESQ for POS-DAE

System	Noise Reduction		Speech Distortion		PESQ	
	-5dB	-10dB	-5dB	-10dB	-5dB	-10dB
Baseline DAE	6.867	8.079	3.541	3.956	1.693	1.478
Proposed POS-DAE	2.451	2.824	5.593	6.394	**2.147**	**2.025**
DAE + Oracle SAD	7.179	8.45	1.654	1.654	**2.588**	2.292
POS-DAE + Oracle SAD	4.687	5.696	3.192	3.365	2.538	**2.432**
Type-I Architecture DAE	6.309	7.589	2.924	3.380	2.018	1.857
Type-II Architecture DAE	6.336	7.571	2.905	3.364	2.022	1.856
Type-I Architecture POS-DAE	3.256	3.570	4.933	5.803	2.071	1.951
Type-II Architecture POS-DAE	3.157	3.548	4.954	5.796	2.071	1.951

Table 3: Performance evaluation of Speech Enhancement using Noise Reduction, Speech Distortion and PESQ

proposed 2-stage architecture with and without the introduced loss function.

4.1. Effect of Proposed Objective Loss Function

Effect of Penalty: We evaluated the performance of the POS-DAE system for different values of penalty p and we observe that higher penalty gives better PESQ in Table 2. This provides initial validation for the proposed loss measure. For subsequent experiments, without optimizing, we use a penalty of 10.

Results: As seen from Table 3-top, we find that the traditional DAE shows promising performance in terms of objective measures through speech distortion and noise reduction. However, there is a degradation in PESQ score. The new objective function shows sub-optimal performance in terms of objective measures compared to the traditional DAE, but more importantly shows significant improvements in PESQ values. This indicates that the traditional DAE and the MSE objective function achieves better objective measures sacrificing the perceptual quality of speech. The difference in the objective measures validates the notion that MSE-DAE introduces over-smoothing and clipping of speech signals, thereby supporting the motivation and effectiveness of the POS-DAE objective function. Moreover, we see that the POS-DAE provides better improvements for lower SNR -10dB signals in terms of PESQ, where we expect the over-smoothing to be more severe.

4.2. Effect of Augmentation of Oracle SAD labels

Table 3-middle shows the improvements achieved by providing oracle speech activity regions to both DAE (baseline) and POS-DAE systems. We see significant improvements in terms of both objective and perceptual evaluation measures for both systems over the baseline. POS-DAE achieves the best noise reduction with the least speech distortion, and best perceptual quality. This becomes more pronounced for lower SNR of -10dB. The results encourage us to use POS-DAE and the 2-stage architecture in conjunction for complementary results.

4.3. Evaluation of Proposed 2-Stage Architecture

The 2-stage architecture replaces the Oracle SAD labels with a SAD-DNN system as described in Section 2.2.

Type I: The results in table 3 are obtained after jointly training the 2 stages. The results show a significant improvement over the baseline for Type I architecture applied to DAE. Further improvements in PESQ values are obtained by using a POS-DAE. We again observe significant improvement for low SNR -10dB compared to the DAE system consistent with the findings from oracle SAD experiments. This stresses the fact that POS-DAE performs better at low SNR environment albeit providing slight improvements for comparatively higher SNR environments over the baseline DAE systems.

Type II: Again, the results for Type II are an improvement over the baseline. However, in our experiments we found that there was no significant improvements provided by Type II over Type I, both in terms of objective measures and PESQ values. We believe this is because the increased parameterization is not being exploited during training due to the already good initialization point and the limited (for such a large DNN) training data.

5. Relation to Prior Work

In [32], a weighted denoising autoencoder (WDA) was proposed by altering the MSE loss function with weights associated with different frequency components of the spectrum. Our system could be used in conjunction with such a loss function to further take advantages of the speech spectrum. In [18], a 2-level architecture was proposed with the first stage trained to predict a time-frequency (T-F) binary mask for noise dominance and speech dominance. In our system, we instead train a SAD system and have the following advantages: (i) there is no need to explicitly estimate the noisy spectra as in a T-F mask system (ii) we do not need to set any manual thresholding, since our second stage is exploiting soft-labels of speech activity, and (iii) we proposed jointly training the two modules to expand relevant and contextual information transfer.

6. Conclusions & Future Direction

We introduced a new loss function designed to take into consideration the perceptual speech quality during DAE training and addressed the challenging problem of speech enhancement at low SNR. The proposed Perception Optimized Speech Denoising system was demonstrated to give better PESQ values than a traditional MSE based DAE systems. The difference in the objective measures indicated that the problem of over-smoothing apparent in MSE-DAE systems was mediated by the POS-DAE system. We also proposed two different architectures, based on jointly training SAD and denoising systems which proved to be better than the traditional DAE systems. Even though the jointly-optimized Type-II architecture provided no significant gains over Type-I, this is likely due to the relatively limited training data for such a big network.

In our system, the penalty factor assigned in POS-DAE systems is a simple positive scalar constant. In the future, POS-DAE could be further extended by using a multi-dimensional penalty, corresponding for instance to different penalties per frequency band. For instance, we could assign a lesser penalty for over-smoothing, higher frequency components that contain less speech information. Further, our system could be used in conjunction with WDA introduced in [32] to add more constraints to the loss function and customize it to the application of speech enhancement.

7. References

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

[2] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 27, no. 2, pp. 113–120, 1979. 1

[3] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'79.*, vol. 4. IEEE, 1979, pp. 208–211. 1

[4] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 32, no. 6, pp. 1109–1121, 1984. 1

[5] R. J. McAulay and M. L. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 28, no. 2, pp. 137–145, 1980. 1

[6] 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. 1

[7] M. Dendrinos, S. Bakamidis, and G. Carayannis, "Speech enhancement from noise: A regenerative approach," *Speech Communication*, vol. 10, no. 1, pp. 45–57, 1991. 1

[8] Y. Ephraim and H. L. Van Trees, "A signal subspace approach for speech enhancement," *Speech and Audio Processing, IEEE Transactions on*, vol. 3, no. 4, pp. 251–266, 1995. 1

[9] G. Kim, Y. Lu, Y. Hu, and P. C. Loizou, "An algorithm that improves speech intelligibility in noise for normal-hearing listeners," *The Journal of the Acoustical Society of America*, vol. 126, no. 3, pp. 1486–1494, 2009. 1

[10] S. Srinivasan, N. Roman, and D. Wang, "Binary and ratio time-frequency masks for robust speech recognition," *Speech Communication*, vol. 48, no. 11, pp. 1486–1501, 2006. 1

[11] E. A. Wan and A. T. Nelson, "Networks for speech enhancement," *Handbook of neural networks for speech processing. Artech House, Boston, USA*, vol. 139, 1999. 1, 3

[12] S. Tamura, "An analysis of a noise reduction neural network," in *Acoustics, Speech, and Signal Processing, 1989. ICASSP-89., 1989 International Conference on*. IEEE, 1989, pp. 2001–2004. 1

[13] S. Tamura and M. Nakamura, "Improvements to the noise reduction neural network," in *Acoustics, Speech, and Signal Processing, 1990. ICASSP-90., 1990 International Conference on*. IEEE, 1990, pp. 825–828. 1

[14] F. Xie and D. C. Van, "A family of mlp based nonlinear spectral estimators for noise reduction," in *Acoustics, Speech, and Signal Processing, 1994. ICASSP-94., 1994 IEEE International Conference on*, vol. 2. IEEE, 1994, pp. II–53. 1, 3

[15] H. B. Sorensen, "A cepstral noise reduction multi-layer neural network," in *Acoustics, Speech, and Signal Processing, 1991. ICASSP-91., 1991 International Conference on*. IEEE, 1991, pp. 933–936. 1

[16] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "A regression approach to speech enhancement based on deep neural networks," *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, vol. 23, no. 1, pp. 7–19, 2015. 1, 2, 3

[17] 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. 1, 2

[18] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "Dynamic noise aware training for speech enhancement based on deep neural networks." in *INTERSPEECH*, 2014, pp. 2670–2674. 1, 2, 3, 4

[19] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder." in *INTERSPEECH*, 2013, pp. 436–440. 1, 3

[20] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "An experimental study on speech enhancement based on deep neural networks," *Signal Processing Letters, IEEE*, vol. 21, no. 1, pp. 65–68, 2014. 1, 3

[21] M. Wollmer, Z. Zhang, F. Weninger, B. Schuller, and G. Rigoll, "Feature enhancement by bidirectional lstm networks for conversational speech recognition in highly non-stationary noise," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 6822–6826. 1

[22] A. L. Maas, Q. V. Le, T. M. O'Neil, O. Vinyals, P. Nguyen, and A. Y. Ng, "Recurrent neural networks for noise reduction in robust asr." in *INTERSPEECH*. Citeseer, 2012. 1

[23] Z. Chen, S. Watanabe, H. Erdogan, and J. Hershey, "Integration of speech enhancement and recognition using long-short term memory recurrent neural network," 2015. 1

[24] H. Siln, E. Hel, J. Nurminen, and M. Gabbouj, "Ways to implement global variance in statistical speech synthesis," in *in Proc. Interspeech*, 2012. 2

[25] J. Du, Q. Wang, T. Gao, Y. Xu, L.-R. Dai, and C.-H. Lee, "Robust speech recognition with speech enhanced deep neural networks." in *INTERSPEECH*, 2014, pp. 616–620. 2

[26] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "Global variance equalization for improving deep neural network based speech enhancement," in *Signal and Information Processing (ChinaSIP), 2014 IEEE China Summit & International Conference on*. IEEE, 2014, pp. 71–75. 2

[27] D. B. Dean, S. Sridharan, R. J. Vogt, and M. W. Mason, "The qut-noise-timit corpus for the evaluation of voice activity detection algorithms," in *Interspeech 2010*, Makuhari Messe International Convention Complex, Makuhari, Japan, September 2010, click on the Related data link below to download the QUT-NOISE-TIMIT corpus dataset (listed under the databases tab) of QUT's SAVIT Research Program. [Online]. Available: http://eprints.qut.edu.au/38144/ 3

[28] 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, 1993. 3

[29] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, Dec. 2011, iEEE Catalog No.: CFP11SRW-USB. 3

[30] D. Povey, X. Zhang, and S. Khudanpur, "Parallel training of dnns with natural gradient and parameter averaging," *arXiv preprint arXiv:1410.7455*, 2014. 3

[31] 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. 3

[32] B. Xia and C. Bao, "Speech enhancement with weighted denoising auto-encoder." in *INTERSPEECH*, 2013, pp. 3444–3448. 4