SPEECH ENHANCEMENT USING MULTIPLE DEEP NEURAL NETWORKS

Pavan Karjol[+]*, Ajay Kumar M*[*]*, and Prasanta Kumar Ghosh*[+]

[+]Electrical Engineering, Indian Institute of Science, Bengaluru 560012, India
[*]Electronics and Communication Engineering, NIT-K, Surathkal 575025, India

ABSTRACT

In this work, we present a variant of multiple deep neural network (DNN) based speech enhancement method. We directly estimate clean speech spectrum as a weighted average of outputs from multiple DNNs. The weights are provided by a gating network. The multiple DNNs and the gating network are trained jointly. The objective function is set as the mean square logarithmic error between the target clean spectrum and the estimated spectrum. We conduct experiments using two and four DNNs using the TIMIT corpus with nine noise types (four seen noises and five unseen noises) taken from the AURORA database at four different signal-to-noise ratios (SNRs). We also compare the proposed method with a single DNN based speech enhancement scheme and existing multiple DNN schemes using segmental SNR, perceptual evaluation of speech quality (PESQ) and short-term objective intelligibility (STOI) as the evaluation metrics. These comparisons show the superiority of proposed method over baseline schemes in both seen and unseen noises. Specifically, we observe an absolute improvement of 0.07 and 0.04 in PESQ measure compared to single DNN when averaged over all noises and SNRs for seen and unseen noise cases respectively.

Index Terms— Deep neural networks, speech enhancement, gating network.

1. INTRODUCTION

Several methods exist for speech enhancement including Wiener filtering [1], perceptually enhanced KLT [2], minimum mean square error (MMSE) estimation [3] and deep neural networks (DNN) [4]. These speech enhancement methods can be grouped into two broad categories, namely, statistical approaches like Wiener filtering [1], minimum mean square estimation (MMSE) [5] etc and data driven methods like neural networks [6]. In statistical methods, particular probabilistic models are assumed for speech and noise. Ephraim et al. [5], proposed a minimum mean square estimation technique of clean speech. It is based on the independent complex Gaussian distribution assumption on speech and noisy spectral coefficients. Assuming that the logarithmic error is perceptually more suitable for speech enhancement, a logarithmic minimum mean square estimation [3] technique was proposed. These statistical algorithms require a running estimate of noise and clean speech variances. However, such estimates are typically poor for highly non-stationary noises. In addition, these algorithms work under the assumption that spectral coefficients are uncorrelated in a speech frame. However, it is well known that spectral coefficients are correlated in different frequencies as well as at different time instants [7].

Over the last decade, the data driven methods have been shown to provide better results than traditional statistical methods. Among these data driven methods, neural networks are the state of the art techniques. Hence neural networks [8–10] are employed for speech enhancement as well. Tamura et al. [8], proposed to use a shallow network to estimate the clean speech with input to the network set as noisy speech. But these algorithms did not provide satisfactory results due to less amount of data and smaller network size. In addition the bigger networks suffered from the problem of getting stuck in local minima. In recent years, a number of modifications (both in training of the neural network, and in architecture) are proposed to alleviate these inherent problems of neural networks. Hinton et al. [11], proposed a greedy layer wise training algorithm. However, later, better random initialization [12] techniques and better activation functions like *relu* [13], have been shown to provide similar (or better) performance. Xu et al. [4], proposed a DNN based speech enhancement technique, where the DNN is initialized with the weights provided by a restricted boltzmann machine (RBM) which is trained layer wise.

There are a number of data driven speech enhancement techniques where multiple experts are used. For example, a mixture maximum model was proposed by Amit et al. [14], which is based on broad phoneme classes. However, it requires prior enhanced Mel frequency cepstral coefficient (MFCC) vectors specific to each phoneme, which makes the algorithm less general across different speakers and also with respect to the intra broad phoneme class variability. A phoneme information based approach using DNNs was proposed by Wang et al. [15]. The algorithm involves training of forty DNNs, one for each phoneme class to estimate the ideal ratio mask (IRM). During testing, an automatic speech recognition (ASR) system is used to predict the correct phoneme label. The DNN corresponding to the predicted phoneme label is used to estimate the IRM. However, this method may result in speech quality degradation due to the poor performance of ASR system in noisy conditions. Chazan et al. [16] employed a similar method where speech presence probability (SPP) is estimated instead of IRM and a classifier network is used in place ASR system.

We hypothesize that the method of clustering the input data based on phoneme specific information during training [15], [16], may not be optimal. This could be because a clustering method different from that using phonetic groups could be a better choice for enhancing speech. Therefore, in our work, we propose an enhancement scheme that doesn't restrict the system to be trained with phonetic information. Moreover, unlike systems that require about forty DNNs [16], in the proposed method of using multiple DNNs, we alleviate the need for using such a complicated network. Multiple networks based techniques for learning have been used in the past. For example, a mixture of experts was proposed by Jacobs et al. [17]. It involves partitioning the training data and learning a separate network for each partition. These two tasks are done in conjunction. However, each expert is limited to a linear regression model. In the proposed work we employ a similar architecture using DNNs for speech enhancement. Although such an architecture is also used by Chazan et al. [18], it differs from the proposed method with regard to what is being estimated by the network. Specifically,

while Chazan et al. [18], use the network to estimate the speech presence probability (SPP), we employ the network to estimate the clean speech spectrum, directly. Thus, our multiple DNN network, involves the joint training of N DNNs in conjunction with a gating network, such that the final estimate of the clean spectrum is a linear combination of the outputs from N DNNs, weighed by the output of the gating network.

We conduct experiments with clean speech utterances taken from the TIMIT corpus, nine noise types taken from AURORA database (four among which are used for training) in four SNR conditions. The performance metrics used are perceptual evaluation of speech quality (PESQ) [19], segmental signal-to-noise ratio (seg SNR) and short-time objective intelligibility (STOI) [20]. We compare the proposed method against single DNN with different number of parameters. We also compare with SPP based multiple DNN system. We observe that the proposed method outperforms these baseline schemes in most cases. Specifically we observe an improvement of 0.07 in PESQ (averaged across all SNRs and seen noises) for seen noises compared to the single DNN case. When averaged across unseen noise cases the corresponding improvement is 0.04 over single DNN case.

2. SPEECH ENHANCEMENT USING MULTIPLE DEEP NEURAL NETWORKS

Speech is composed of a rich variety of spectro-temporal structures. These structures largely vary with a number of factors including different types of sounds, i.e., phonemes and speakers. A single DNN system may not capture these varying structures completely, thus giving rise to the need for employing multiple DNNs [16] [15]. As shown in Fig 1, the multiple DNN based system involves N DNNs, each contributing to the final enhanced speech, and a gating network which provides the weights to combine the outputs of the N DNNs. In general the final output can be obtained as,

$$\hat{y} = \sum_{k=1}^{N} p_k(x) f_k(x), \tag{1}$$

where $f_k(x)$ is the output of k^{th} DNN and $p_k(x)$ is the corresponding weight for the k^{th} output for the input x. Single DNN is a trivial case corresponding to $N = 1$ and $p_1(x) = 1, \forall x$.

There are three major aspects involved in a multiple DNN system for speech enhancement: 1) the number DNNs to be chosen, 2) the target output representation, 3) the objective function used to train the network.

Chazan et al. [16], used forty DNNs (with all having same architecture) corresponding to each phoneme. However, such a network typically overfits the data, since there won't be enough data for certain phonemes compared to other ones. In addition, due to high number of classes, (i.e., the number of phonemes) the error on cross phonemes (i.e., when a DNN trained on a particular phoneme encounters other phonemes during testing) will play a significant role. This, in turn, affects the overall performance. Therefore, using a large number of DNNs may not be ideal in such a scenario. Instead we can use less number of DNNs so that each DNN can be trained well with reasonable amount of data. In this regard, Chazan et al. [18], used lesser number of DNNs and trained the entire system jointly. In other words, all the N DNNs and the gating network are trained jointly in a completely data driven way. The objective function used in this case is given by,

$$E_r = \sum_{i=1}^{M} d\Big(y_i, \sum_{k=1}^{N} p_k(x_i) f_k(x_i)\Big), \tag{2}$$

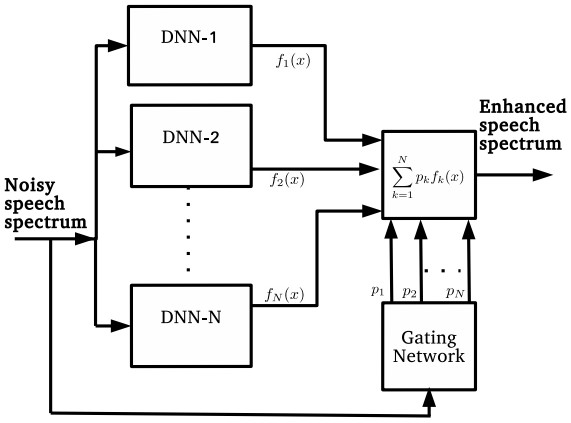

Fig. 1: Multiple DNN based speech enhancement system

where $d()$ is the error metric used for training. x_i and y_i are i^{th} training example for noisy speech input, clean speech target output respectively. $\hat{y_i}$ is the estimated clean speech. However, the target output in the work by Chazan et al. [18], was set as SPP. In this work we directly try to estimate clean speech spectrum from such a network and experimentally demonstrate that the use of clean speech spectrum as the output representation results in a better speech enhancement performance. In the context of using an objective function, Jacobs et al. [17], proposed to use the following objective function,

$$E_r = -\sum_{i=1}^{M} \log\left(\sum_{k=1}^{N} p_k(x_i) \exp\big(-\frac{1}{2} \parallel y_i - f_k(x_i) \parallel\big)\right), \tag{3}$$

which could enable each DNN to perform well on certain region of input space. In the current work, in addition to using eq. (2), we also experiment with eq. (3) as objective function for training and then examine which of these objective functions, when used for training the multiple DNN system, results in a better enhancement performance.

3. EXPERIMENTS

3.1. Database

Clean speech utterances are taken from TIMIT [21]. It consists 4620 recordings in training set and 1680 recordings in test set. The sampling frequency of these recordings is 16 kHz. The noise signals used for the experiments are white, babble, restaurant, street, airport, car, exhibition, subway and train. Except white noise all noise files are taken from AURORA [22] database. White noise (white gaussian noise) is synthetically generated. The sampling rate of the noise signals is 8 kHz. To match the sampling frequencies, we downsample clean speech utterances of TIMIT to 8 kHz.

3.2. Experimental setup

The noise signals used for training (seen noises) are white, babble, restaurant and street. We add noise signals to clean speech at global SNRs of -5, 0, 5, and 10dB. The frame length is set to be 256 with 50% overlap. The TIMIT utterances are added with training noises at various SNRs. For every noise and SNR combinations, 100k frames are randomly picked from the entire noisy TIMIT training set. This results in a total of 1600k frames (100k frames $\times$ 4 noises $\times$ 4 SNRs). 80% of these frames are used for training and remaining 20% are used as the validation set. For testing we use 250

TIMIT test utterances added with nine noises (i.e., four seen training noises and five unseen noises) at four different SNRs.

Proposed method: We use four (N=4) DNNs and a gating network, each of which are 3 layer deep. The number of units at each layer is set to be 512. This whole network with target output set as clean speech spectrum is referred to as M-DNN$_{P4}$. The activation at each layer is set to be *relu*. We also test the proposed method using two DNNs and a gating network with same number of hidden layers as that of M-DNN$_{P4}$. We also examine the proposed system with $N = 2$. We consider two such cases based on the number of units at each layer viz 1024 and 512 units. These systems are referred to as M-DNN-1$_{P2}$ and M-DNN-2$_{P2}$ respectively. The multiple DNN system with four DNNs and 1024 units at each layer, is found to overfit the data. Hence, we do not report results with such a network. The objective function used for training is given by eq. (2).

Single DNNs: For the single DNN baseline, we use a 3 layered architecture with equal number of units at each layer. We have two variants of these DNNs: 1) S-DNN$_1$ which has 1024 units at each layer and 2) S-DNN$_2$ which has 1324 units at each layer. The number of units at each layer in S-DNN$_2$ is chosen such that, it results in a number of parameters identical to that of M-DNN$_{P4}$. The activation at each layer is set to be *relu*. Single DNN with number of parameters equated to that M-DNN-1$_{P2}$ is found to overfit the data. Hence, the corresponding results are not reported.

SPP based speech enhancement: Chazan et al. [18], proposed to use a similar network for speech enhancement except that the target output is set as SPP. The clean speech log spectrum $\hat{y}$ is estimated using the predicted SPP as follows,

$$\hat{y} = x - \beta(1 - \rho), \tag{4}$$

where β is the attenuation constant, ρ is the predicted SPP and x is the noisy speech log spectrum. For this approach, the network architecture is same as that of M-DNN$_{P4}$ except the target output is set as SPP. In addition, the output activation is set to be *sigmoid* (as proposed in [18]) instead of *relu*. This system is referred as M-DNN$_{S4}$.

Speech enhancement with competitive learning: We also experiment with the objective function given in eq. (3). The network architecture is kept as that of M-DNN$_{P4}$ except the objective function to train the network.

We experiment with architectures for the SPP and competitive learning methods, similar to those of M-DNN-1$_{P2}$, M-DNN-2$_{P2}$ and M-DNN$_{P4}$. Among these, we observe that an architecture similar to that of M-DNN$_{P4}$ performs the best for both these cases, and hence, report the results only for this variant.

3.3. DNN training parameters

We train every system for 50 epochs with early stopping criteria [23]. The optimizer used is *adam* with default parameters [24]. The input to each system is normalized speech spectrum with context length of two frames (previous two and next two frames). For the proposed method and single DNN schemes, the error metric used is 'mean square logarithmic error'(*msle*) [23], while for SPP based approach the error metric used is 'mean square error' (*mse*) (as proposed in [18]).

3.4. Evaluation metrics

We evaluate the different schemes using PESQ, seg SNR and STOI. PESQ is a measure of perceptual quality of speech, while STOI measures the intelligibility. seg SNR provides information about average reconstruction error across frames with respect to the clean speech. Hence, these measures are used to objectively evaluate the enhanced speech.

Enhancement Scheme	i/p SNR			
	-5 dB	0 dB	5 dB	10 dB
M-DNN$_{S4}$	1.7517	2.1166	2.4715	2.8141
M-DNN$_{C4}$	2.1358	2.4331	2.6627	2.8455
S-DNN$_1$	2.2209	2.4677	2.6603	2.7986
S-DNN$_2$	2.1915	2.4685	2.6712	2.8061
M-DNN-1$_{P2}$	2.2051	2.4714	2.6721	2.8074
M-DNN-2$_{P2}$	2.2321	2.5137	2.7479	2.9323
M-DNN$_{P4}$	**2.2709**	**2.5582**	**2.7959**	**2.9886**

Table 1: Average PESQ results for seen noise cases

Enhancement Scheme	i/p SNR			
	-5 dB	0 dB	5 dB	10 dB
M-DNN$_{S4}$	-3.1789	-0.7345	1.8513	4.2052
M-DNN$_{C4}$	0.0056	1.7270	3.4493	5.0656
S-DNN$_1$	0.4401	2.0055	3.6300	5.1630
S-DNN$_2$	0.4247	2.0246	3.6016	5.0035
M-DNN-1$_{P2}$	0.4650	2.0495	3.6533	5.1094
M-DNN-2$_{P2}$	0.3887	1.9920	3.7289	5.4650
M-DNN$_{P4}$	**0.5695**	**2.2634**	**4.0867**	**5.8773**

Table 2: Average seg SNR results for seen noise cases

Enhancement Scheme	i/p SNR			
	-5 dB	0 dB	5 dB	10 dB
M-DNN$_{S4}$	0.6135	0.7347	0.8301	0.8920
M-DNN$_{C4}$	0.7222	0.8042	0.8616	0.8998
S-DNN$_1$	0.7381	0.8114	0.8627	0.8962
S-DNN$_2$	0.7374	0.8103	0.8603	0.8925
M-DNN-1$_{P2}$	0.7348	0.8092	0.8608	0.8939
M-DNN-2$_{P2}$	0.7395	0.8176	0.8725	0.9089
M-DNN$_{P4}$	**0.7441**	**0.8224**	**0.8777**	**0.9144**

Table 3: Average STOI results for seen noise cases

Enhancement Scheme	i/p SNR			
	-5 dB	0 dB	5 dB	10 dB
M-DNN$_{S4}$	1.6939	2.0347	2.3849	**2.7391**
M-DNN$_{C4}$	1.7530	2.0718	2.3814	2.6730
S-DNN$_1$	1.7285	2.0468	2.3504	2.6198
S-DNN$_2$	1.7348	2.0642	2.3719	2.6348
M-DNN-1$_{P2}$	**1.7540**	2.0773	2.3860	2.6486
M-DNN-2$_{P2}$	1.7324	**2.0854**	**2.4128**	2.7230
M-DNN$_{P4}$	1.6005	2.0334	2.4002	2.7324

Table 4: Average PESQ results for unseen noise cases

4. RESULTS AND DISCUSSION

The average values of seg SNR, PESQ and STOI for seen noise cases are presented in Tables (1 - 3). Similarly the average results for unseen noise cases are given in Tables (4 - 6). We observe that the proposed method outperforms the baseline schemes in most cases. In few cases viz for unseen noises in low SNR conditions, using multiple DNN with direct clean spectrum estimation doesn't yield better results compared to the single DNN schemes. This may be due to the fact that, at low SNRs (for unseen case), the structure in the spectrum is less distinguishable across different clusters learnt

Enhancement Scheme	i/p SNR			
	-5 dB	0 dB	5 dB	10 dB
M-DNN$_{S4}$	-3.8267	-1.5364	1.1804	3.7280
M-DNN$_{C4}$	**-2.8976**	-0.4222	2.0844	4.3052
S-DNN$_1$	-3.2850	-0.7160	1.9376	4.3077
S-DNN$_2$	-3.0418	-0.3964	2.1724	4.3191
M-DNN-1$_{P2}$	-3.0239	**-0.3905**	2.2362	4.4494
M-DNN-2$_{P2}$	-3.0849	-0.4768	2.2413	4.7173
M-DNN$_{P4}$	-3.5594	-0.6332	**2.2509**	**4.8757**

Table 5: Average seg SNR results for unseen noise cases

Enhancement Scheme	i/p SNR			
	-5 dB	0 dB	5 dB	10 dB
M-DNN$_{S4}$	0.5756	0.7087	0.8180	0.8885
M-DNN$_{C4}$	**0.5851**	0.7224	0.8272	0.8921
S-DNN$_1$	0.5696	0.7117	0.8192	0.8849
S-DNN$_2$	0.5775	0.7195	0.8225	0.8830
M-DNN-1$_{P2}$	0.5832	**0.7237**	0.8263	0.8858
M-DNN-2$_{P2}$	0.5746	0.7201	**0.8302**	0.8976
M-DNN$_{P4}$	0.5294	0.7092	0.8296	**0.9002**

Table 6: Average STOI results for unseen noise cases

Enhancement Scheme	noise type (seen or unseen)	
	seen	unseen
M-DNN$_{S4}$	2.2885	2.2131
M-DNN$_{C4}$	2.5193	2.2198
S-DNN$_1$	2.5369	2.1864
S-DNN$_2$	2.5343	2.2014
M-DNN-1$_{P2}$	2.5390	2.2165
M-DNN-2$_{P2}$	2.6065	**2.2384**
M-DNN$_{P4}$	**2.6534**	2.1916

Table 7: PESQ results averaged over SNRs and noises

after training the multiple DNN network due to high amount of noise (which itself is unseen).

4.1. Dependency on the type of target output

From the results provided in Tables (1-6), we note that the proposed schemes (M-DNN-1$_{P2}$, M-DNN-2$_{P2}$ and M-DNN$_{P4}$) perform better than SPP based ones (M-DNN$_{S4}$). This suggests that direct estimation of clean speech spectrum using multiple DNN is better than estimating SPP using similar network. It could be due to the fact that the approximation for log of sum of two spectra ($\log S = \log(S1 + S2)$) used in SPP based method [16] i.e., $log(S1 + S2) \simeq \max(\log(S1), \log(S2))$ does not hold well at low SNRs, unlike that at high SNRs. In addition, the clean speech is estimated using the eq. (4). As we notice, the clean speech estimation depends on an attenuation constant β [16] unlike any such in the proposed approach. The value of this constant is kept fixed while its optimal value may vary with the SNR.

4.2. Dependency on the objective function

From the comparisons presented in Tables (1-6), we notice that M-DNN$_{P4}$ performs better than M-DNN$_{C4}$ in most of the cases except at -5 dB SNR for unseen noise cases. Although the objective function in eq. (3) enables competitive learning of the individual DNN, the overall performance in terms of reconstruction error of clean speech spectrum is not better than the network trained with the objective function in eq. (2). We also note that the weighted average of outputs from different DNNs behaves in a way identical to that of residual networks [25]. We observe that, only one DNN contributes to the overall output prominently. The other DNNs model the residual error. This is illustrated in Fig 2 in terms of seg SNR measure averaged across 250 test sentences for all seen noise and SNR combinations (a total of 250*4*4 sentences) with standard deviation shown in red errorbar. We illustrate with the seg SNR measure since it correlates well (compared to PESQ) with the reconstruction error. Statistical tests show that the average seg SNR using individual DNNs are significantly different. The DNNs in the M-DNN$_{P4}$ system contribute to the final output in the following order: DNN$_4$ > DNN$_3$ > DNN$_1$ > DNN$_2$ except at the i/p SNR of -5 dB. This indicates the similarity of the multiple DNN system with the residual network.

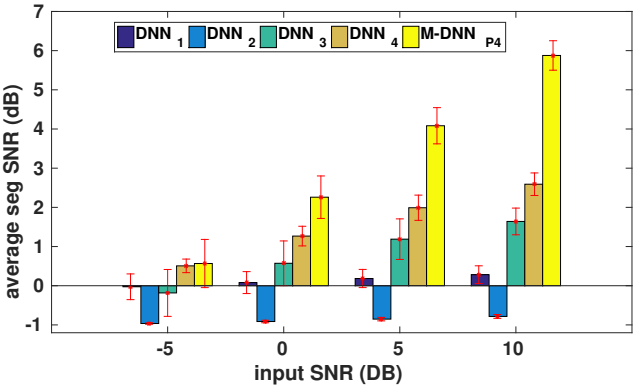

Fig. 2: Comparison of individual DNN performances in the M-DNN$_{P4}$ system. DNN$_k$ corresponds to the k^{th} DNN in the M-DNN$_{P4}$ system. seg SNR values are averaged over training noise cases. The DNN$_4$ contributes prominently to the overall seg SNR.

4.3. Dependency on the number of DNNs

We observe a trade off between the performances for seen noise cases and for unseen noise cases with respect to the number of DNNs used. As we increase the number of DNNs, the performance in seen noise cases at high SNRs improves at the cost of the deterioration at the remaining cases.

In order to obtain an overall performance comparison among different schemes, the PESQ values are averaged over all SNRs separately for seen and unseen noise cases. Table 7 shows the average PESQ values. It is clear that the proposed multiple DNN network with clean spectrum as the output representation performs better than all other schemes for both seen and unseen noise cases. This demonstrates the benefit of the proposed multiple DNN based speech enhancement scheme.

5. CONCLUSION

In this work, we propose to use speech spectrum as the output representation in a multiple DNN based speech enhancement scheme. In order to show the benefit of the proposed scheme, we compare different schemes, by varying objective functions, and target outputs. We observe that, weighted average of outputs from different DNN, with target output set as clean speech spectrum, gives better performance compared to other schemes for both seen and unseen noises considered in this work. In future work, we plan to put more structure on the weights used to compute the output. Optimization of architecture of each DNN could also provide room for further improvement.

6. REFERENCES

[1] K. Paliwal and A. Basu, "A speech enhancement method based on kalman filtering," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'87.*, vol. 12. IEEE, 1987, pp. 177–180.

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

[3] Y. Ephraim and D. Malah, "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.

[4] 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, Jan 2014.

[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] B. Yegnanarayana, *Artificial neural networks*. PHI Learning Pvt. Ltd., 2009.

[7] I. Cohen and S. Gannot, "Spectral enhancement methods," *Springer*, pp. 873–902, 2008.

[8] S. Tamura, "An analysis of a noise reduction neural network," *International Conference on Acoustics, Speech, and Signal Processing,*, vol. 3, pp. 2001–2004, May 1989.

[9] F. Xie and D. V. Compernolle, "A family of MLP based nonlinear spectral estimators for noise reduction," *International Conference on Acoustics, Speech, and Signal Processing,*, vol. 2, pp. II/53–II/56, Apr 1994.

[10] 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.

[11] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[12] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks." in *Aistats*, vol. 9, 2010, pp. 249–256.

[13] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th international conference on machine learning (ICML-10)*, 2010, pp. 807–814.

[14] A. Das and J. H. L. Hansen, "Phoneme selective speech enhancement using parametric estimators and the mixture maximum model: A unifying approach," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 8, pp. 2265–2279, Oct 2012.

[15] Z. Q. Wang, Y. Zhao, and D. Wang, "Phoneme-specific speech separation," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 146–150, March 2016.

[16] S. E. Chazan, S. Gannot, and J. Goldberger, "A phoneme-based pre-training approach for deep neural network with application to speech enhancement," *IEEE International Workshop on Acoustic Signal Enhancement (IWAENC)*, pp. 1–5, Sept 2016.

[17] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton, "Adaptive mixtures of local experts," *Neural Comput.*, vol. 3, no. 1, pp. 79–87, Mar. 1991. [Online]. Available: http://dx.doi.org/10.1162/neco.1991.3.1.79

[18] S. E. Chazan, J. Goldberger, and S. Gannot, "Speech enhancement using a deep mixture of experts," *CoRR*, vol. abs/1703.09302, 2017. [Online]. Available: http://arxiv.org/abs/1703.09302

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

[20] 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.

[21] J. S. Garofolo *et al.*, "Getting started with the DARPA TIMIT CD-ROM: An acoustic phonetic continuous speech database," *National Institute of Standards and Technology (NIST), Gaithersburgh, MD*, vol. 107, 1988.

[22] H. G. Hirsch and D. Pearce, "The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions," *Automatic Speech Recognition: Challenges for the new Millenium ISCA Tutorial and Research Workshop (ITRW)*, 2000.

[23] F. Chollet, "Keras," https://github.com/fchollet/keras, 2017.

[24] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2014. [Online]. Available: http://arxiv.org/abs/1412.6980

[25] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *arXiv preprint arXiv:1512.03385*, 2015.