DNN Based Speech Enhancement for Unseen Noises Using Monte Carlo Dropout

Nazreen P.M.
MILE Lab
Department of Electrical Engineering
Indian Institute of Science
Bangalore 560012
nazreenp@iisc.ac.in

A. G. Ramakrishnan
MILE Lab
Department of Electrical Engineering
Indian Institute of Science
Bangalore 560012
agr@iisc.ac.in

***Abstract*—In this work, we propose the use of dropout as a Bayesian estimator for increasing the generalizability of a deep neural network (DNN) for speech enhancement. By using Monte Carlo (MC) dropout, we explore whether the DNN can accomplish better enhancement in unseen noisy conditions. Two DNNs are trained on speech corrupted with five different noises at three SNRs, one using conventional dropout and other with MC dropout and tested on speech with unseen noises. Speech samples are obtained from the TIMIT database and noises from NOISEX-92. In another experiment, we train five DNN models separately on speech corrupted with five different noises, at three SNRs. The model precision estimated using MC dropout is used as a proxy for squared error to dynamically select the best of the DNN models based on their performance on each frame of test data. The first set of experiments aims at improving the performance of an existing DNN with conventional dropout for unseen noises, by replacing the conventional dropout with MC dropout. The second set of experiments aims at finding an optimal way of choosing the best DNN model for de-noising when multiple noise-specific DNN models are available, for unseen noisy conditions.**

***Index Terms*—speech enhancement, deep neural networks, DNN, dropout, unseen noise, Monte Carlo, model uncertainty.**

I. Introduction

Single channel speech enhancement has been a challenging problem for decades. Speech enhancement techniques find several applications such as automatic speech recognition, hearing aids and speaker recognition. Methods proposed in the past include unsupervised methods such as spectral subtraction [1], [2], Wiener filtering [3], minimum mean-square error estimators [4], estimators based on Gaussian prior distributions [5], [6] and residual-weighting schemes [7]–[9]. Most of these methods may perform poorly when the background noise is non-stationary and in unexpected acoustic conditions.

In supervised learning methods, prior information is fed into the models and hence they are expected to perform better than unsupervised methods [10]–[12]. Neural networks have been shown to learn the mapping between noisy and clean speech [13]–[15]. However, these models are small networks with a single hidden layer and cannot fully learn the mapping. Deep architectures have conquered this area recently, since these networks with multiple layers have been shown to better learn the complex mapping between noisy and clean features and hence give really good enhancement performances. Hinton *et al.* proposed a greedy, layer-wise unsupervised learning algorithm [16], [17]. Mass et al. [18] use deep recurrent neural networks for feature enhancement for noise robust automatic speech recognizer (ASR) .

One of the major issues encountered by deep neural network (DNN) based enhancement is the degradation of performance for noises unseen during training. The model learns mapping between noisy and clean speech well for those noises with which it is trained, but performs poorly on speech corrupted by an unseen noise. In fact, this itself could be dealt with as a challenging task in speech enhancement scenario. Though not dealt with separately, techniques have been proposed in the past to address this problem. In [19], they have proposed a regression DNN-based speech enhancement framework, where they train a wide neural network using a good collection of data of about 100 hours of various noise types. In [20], a DNN-SVM (support vector machine) based system is trained on a variety of acoustic data for a huge amount of time. A noise-aware training technique is adopted in [21], where a noise estimate is appended to the input feature for training. They use about 2500 hours of data for training the network.

Hinton [22], [23] introduced the concept of dropout to reduce overfitting during DNN training. Though dropout omits weights during training, it is inactive during the inference stage, whereby all the neurons contribute to the prediction.

Gal and Ghahramani [24] proposed the use of dropout during testing, by showing a theoretical relationship between dropout and approximate inference in a Gaussian process. In [25], they show that by enabling dropout during testing, and averaging the results of multiple stochastic forward passes, the predictions usually become better. They refer to this technique as Monte Carlo (MC) dropout, where the output samples are MC samples from the posterior distribution of models. In [24], they show that the model uncertainty can also be estimated from these samples.

In this work, we explore the use of MC dropout to improve the generalizability of speech models, thereby improving the enhancement performance in a mismatched condition. We show that when the input is a noisy speech corrupted with an unseen noise, the use of MC dropout instead of normal dropout

[22] [23] may give a better output. Hence the same concept could be applied to any of the above mentioned DNN speech models to further improve the generalizability of the output to get a better performance during unseen noise scenarios. We show that using MC dropout has some promise in improving the enhancement performance for unseen noisy scenarios.

We also explore the usage of model uncertainty in problems where multiple noise specific DNN models are used. In a general scenario, one needs to identify the noise type to choose the right noise model to enhance the input noisy speech. However, in the case of an unseen noise scenario, the selection of the appropriate model becomes tricky. By using model uncertainty as an estimate of the prediction error for a sample, this technique can enable the selection of the model with the least prediction error on a frame by frame basis. A similar approach of selecting the best model based on an error estimate is proposed in [26] for robust SNR estimation. They trained a separate DNN as a classifier to select a particular regression model for SNR estimation. However, this approach does not ameliorate the original problem of mismatch in training and testing conditions. In this proposed algorithm, we use the intrinsic uncertainty of a model to estimate the prediction error. Since this method extracts information from the model itself, it has the potential to be a better representative of the prediction error. Our method also circumvents the issue of unseen testing conditions, since according to [24], the model uncertainty itself is an indicator of unseen data. The experiments show that the stronger the correlation between the model uncertainty and the squared error, the better is the enhancement performance.

The aim of our first experiment is to improve the generalizability of an existing DNN denoiser, in the case of unseen noises, by replacing the conventional dropout with MC dropout. The second experiment aims at finding an optimal way of selecting one out of many noise specific DNN models, for unseen noises, in the scenario where multiple models are available for enhancement.

II. DNN BASED SPEECH ENHANCEMENT

Under additive model, the noisy speech can be represented as,

$$x_t(m) = s_t(m) + n_t(m) \tag{1}$$

where $x_t(m)$, $s_t(m)$ and $n_t(m)$ are the m^{th} samples of the noisy speech, clean speech and noise signal, respectively. Taking the short time Fourier transform (STFT), we have,

$$x(\omega_k) = s(\omega_k) + n(\omega_k) \tag{2}$$

where $\omega_k = (2\pi k/R)$, $k = 0, 1, 2...R-1$, k is the index and R is the number of frequency bins. Taking the magnitude of the STFT, the noisy speech can be approximated as

$$X \approx S + N \in \mathbb{R}^{R\times 1} \tag{3}$$

where S and N represent the spectra of the clean speech and the noise, respectively.

A DNN based regression model is trained using the magnitude STFT features of noisy and clean speech, respectively as input and output. The noisy features are then fed to this trained DNN to predict the enhanced features, $\hat{S}$. The enhanced speech signal is obtained by using the inverse Fourier transform of $\hat{S}$ with the phase of the noisy speech signal and overlap-add method.

A. Basic DNN architecture

The proposed baseline system uses a DNN to learn the complex mapping of input noisy speech to clean speech. It consists of 3 fully connected layers of 2048 neurons and an output layer with 257 neurons. We use ReLU non-linearity as the activation function in all the three layers. Our output activation is also ReLU to account for the nonnegative nature of STFT magnitude. Stochastic gradient descent is used to minimize the mean square logarithmic error (E_r) between the noisy and clean magnitude spectra:

$$E_r = \frac{1}{R}\sum_{k=1}^{R}(log(S(k)+1) - log(\hat{S(k)}+1))^2 \tag{4}$$

where $\hat{S}$ and S denote the estimated and reference spectral features, respectively, at sample index k. This baseline system is trained on speech corrupted with five noises and 3 different SNRs using conventional dropout [22] [23].

III. PROPOSED METHODS FOR GENERALIZED SPEECH MODELS

Gal and Ghahramani [24] have shown a theoretical relationship between dropout [22] and approximate inference in a Gaussian process. The proposed system augments the baseline system by dropout as a Bayesian approximation. By using this approximation, a distribution over the weights is learnt, thereby giving uncertainty of the output.

The network output is simulated with input X, using dropout same as that employed during the training time. During testing, T repetitions are performed, with random units in the network dropped out every time, obtaining the results $\{\hat{S_t(X)}\}; 1 \leq t \leq T$. It is shown in [24] that averaging forward passes through the network is equivalent to Monte Carlo integration over a Gaussian process posterior approximation. Empirical estimators of the predictive mean ($E(S)$) and variance (uncertainty, $Var(S)$) from these samples are given as:

$$E(S) \approx \frac{1}{T}\sum_{t=1}^{T}\hat{S_t(X)} \tag{5}$$

$$Var(S) \approx \tau^{-1}I_D + \frac{1}{T}\sum_{t=1}^{T}\hat{S_t(X)}^T\hat{S_t(X)} - E(S)^T E(S) \tag{6}$$

where $\tau = l^2p/2N\lambda$; l: defined prior length scale, p: probability of the units not being dropped, N: total input samples, λ: regularisation weight decay, which is zero for our experiments.

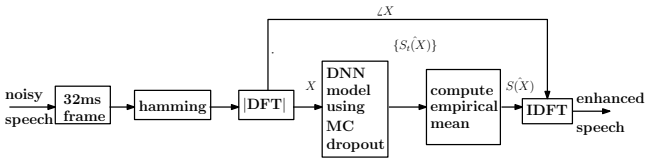

Fig. 1. Enhancement using a DNN-MC dropout model trained on multiple noises.

A. Single DNN-MC dropout model(MC) trained on five noises

A single DNN model is trained using MC dropout with speech corrupted with various noises and SNRs. Figure 1 shows the block diagram of the proposed approach. During testing of MC dropout model, given a noisy speech frame X, multiple repetitions are performed by dropping out random units each time giving T different outputs, $\{S_t(\hat{X})\}; 1 \leq t \leq T$. The empirical mean of these outputs is used as the estimated output $S(\hat{X})$ (5). Enhanced speech is obtained as the inverse Fourier transform of $S(\hat{X})$ with the phase of the noisy speech signal and overlap-add method.

B. Predictive variance (model uncertainty) as the selection criterion to choose one out of many noise specific DNN models using MC dropout (MC-Var)

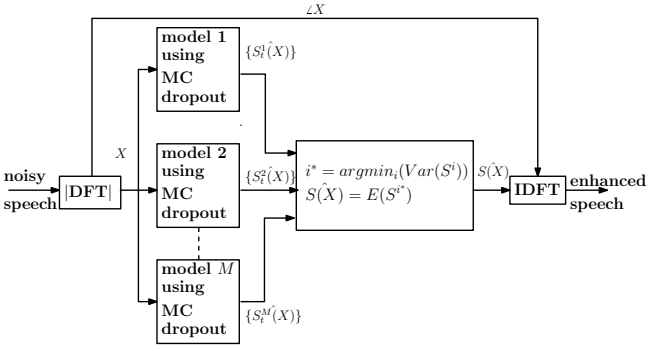

Fig. 2. Enhancement with predictive variance as the criterion to select one out of many noise-specific DNN-MC dropout models

Model-specific enhancement techniques depend on a model selector [27]–[29], which ensures that the model chosen for enhancing each frame entails an overall improved performance. Given multiple noise-specific DNN models for enhancing a frame of noisy speech, one method to select the appropriate model is to detect the type of noise. However, if speech has been corrupted with an unseen noise, the selection of the appropriate model gets harder since the noise detector assumes that one of the models is trained with the correct noise.

In this work, we follow [24] and say that since model uncertainty gives the intrinsic uncertainty of the model for a particular input, we can use it as an estimate of model error. Using this, we can build a framework as per Fig. 2 to enhance speech. Thus, this approach works only when there is a strong correlation between model uncertainty and output error.

Figure 2 shows the block diagram of the proposed approach. Each of the M DNN-MC dropout models is trained on speech corrupted with a particular noise at various SNRs. The architecture is the same as the one defined in Sec. II-A. For a given noisy input frame X, each of these models generates an output by dropping out random units. T repetitions are performed by each model by dropping different units every time, obtaining results $\{S_t^i(\hat{X})\}; 1 \leq t \leq T; 1 \leq i \leq M$; where i is the model index and $M = 5$. The predictive variance (uncertainty) (6) is computed for each of the M different results. The model with the minimum variance is selected as the best one for that frame. The enhanced output $\hat{S}$ is estimated as the empirical mean of the T outputs: $\{S_t^{i*}(\hat{X})\}; 1 \leq t \leq T$. The enhanced speech signal is obtained as the inverse Fourier transform of $\hat{S}$ with the phase of the noisy speech signal and overlap-add method.

IV. Experiments and Results

A. Experimental setup

TIMIT [30] speech corpus is used for the study. The noises used from the NOISEX-92 [31] database are downsampled to 16 kHz to match the sampling rate of TIMIT, in order to synthesize noisy training and test speech signals. The magnitude STFT is computed on frames of size 32 ms with 10 ms frame shift, after applying Hamming window. A 512-point FFT is taken and we use only the first 257 points as input to the DNN, because of symmetry in the spectrum. A DNN based regression model is trained using the magnitude STFT features of clean and noisy speech. For multi-model MC dropout experiments, each DNN model is trained on one of the following noises: Factory 2, m109, leopard, babble and volvo each at SNRs 0, 5 and 10 dB. Thus for our experiments $M = 5$. For the single model case, the DNN is trained on all the above five noises at SNRs 0, 5 and 10 dB for both baseline and MC dropout. During testing, the noisy features are fed to this trained DNN to predict the enhanced features, $\hat{S}$. The enhanced speech signal is obtained as the inverse Fourier transform of $\hat{S}$, using the phase of the noisy speech signal and overlap-add method.

The DNN architecture used has been defined in Sec. II-A. For our experiments, the number of repetitions T is chosen as 50. The Adam optimizer [32] is chosen, whose default regularization weight decay, λ is zero and thus, $\tau^{-1} = 0$ in (6).

B. Results and discussion

Table I lists the improvements obtained in terms of sum squared error (SSE), and segmental SNR (SSNR) [33] for single DNN-MC dropout model (MC) over the conventional DNN dropout model as the baseline for three unseen noises. The values given under the column 'noisy input' show the SSE and SSNR values of the noisy signal input to the DNN models. We use white, pink and factory 1 noises as unseen noises. The reported results are the average over 50 files randomly selected from TIMIT [30] test set. The model achieves a better performance in most of the cases. Though the improvement is not much in terms of SSNR, the SSE values show promise. It is to be noted that the improvement is significant in terms of

TABLE I

PERFORMANCES OF SINGLE DNN-MC DROPOUT MODEL (MC) AND MULTIPLE DNN-MC DROPOUT MODELS WITH PREDICTIVE VARIANCE BASED SELECTION (MC-VAR) ON *unseen* NOISES IN TERMS OF SUM SQUARED ERROR (SSE) AND SEGMENTAL SNR (SSNR). FACTORY2, M109, LEOPARD, BABBLE AND VOLVO NOISES AT SNRS OF 0, 5 AND 10 *dB* ARE USED TO TRAIN THE MODELS

		White				Pink				Factory1			
SNR (dB)	**Metric**	**Noisy input**	**Baseline**	**MC**	**MC-Var**	**Noisy input**	**Baseline**	**MC**	**MC-Var**	**Noisy input**	**Baseline**	**MC**	**MC-Var**
-10	**SSE x10^4**	3.64	3.36	**3.14**	**3.2**	3.96	0.874	**0.848**	**0.708**	3.69	0.720	**0.70**	**0.677**
	SSNR	-8.9	-8.5	**-8.4**	**-8.4**	-8.8	-6.7	**-6.6**	**-5.4**	-8.7	-6.0	**-5.9**	**-5.3**
-5	**SSE x10^4**	1.12	0.960	**0.913**	**0.936**	1.22	0.270	**0.251**	**0.261**	1.12	0.213	**0.200**	**0.20**
	SSNR	-7.2	-6.6	**-6.5**	**-6.5**	-7.1	-4.3	**-4.2**	**-3.7**	-6.9	-3.5	-3.5	**-3.3**
0	**SSE x10^3**	3.41	2.81	**2.60**	**2.70**	3.71	0.858	**0.843**	0.943	3.41	0.682	**0.671**	0.771
	SSNR	-4.6	-3.9	**-3.8**	**-3.8**	-4.5	-1.5	**-1.4**	**-1.3**	-4.4	-0.7	-0.7	-0.8
5	**SSE x10^3**	1.03	0.844	**0.827**	0.857	1.12	0.291	**0.288**	0.391	1.02	0.244	**0.242**	0.285
	SSNR	-1.6	-0.7	-0.7	-0.7	-1.4	1.7	1.7	1.6	-1.3	2.2	2.2	2.0
10	**SSE x10^2**	3.08	2.70	**2.67**	2.73	3.41	1.18	**1.16**	1.40	3.09	1.07	**1.06**	1.24
	SSNR	2.0	2.7	2.7	2.7	2.2	4.7	4.7	4.5	2.3	5.0	5.0	4.8

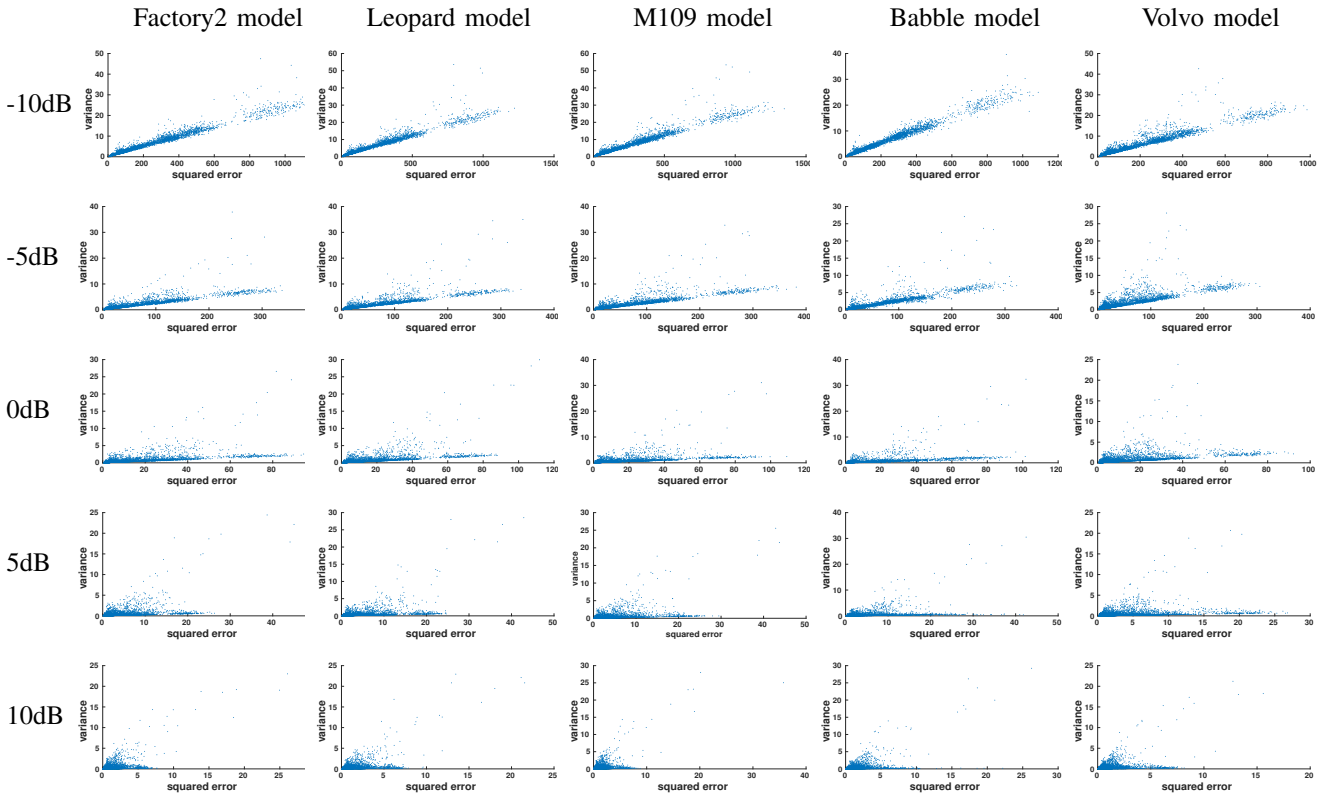

Fig. 3. Correlation plot between the predictive variance and the squared error of the estimated output frames for five different noise-specific DNN models with MC-dropout for the case of speech corrupted with white noise as input

SSE for unseen noises like white noise, especially at low SNRs of -10 and -5 dB. Interestingly, the improvement is negligible or absent with higher SNRs, though the model continues to perform better than the baseline in terms of SSE. Table II shows the performance of the method on seen factory2 noise at SNRs varying from -10 to 10 dB. Though the proposed method does not result in significant improvement on the seen noise, the performance is comparable to the baseline model. Hence, the observations indicate that the proposed method of using MC dropout has the potential to improve generalization performance on unseen noises.

Table I also lists the performance improvements obtained by the multi-model MC dropout DNNs using predictive variance (MC-Var), over the baseline single model with conventional dropout in terms of SSE, and SSNR for speech corrupted with unseen noises of white, pink and factory1, averaged over 50 files randomly selected from TIMIT [30] test set. Again, the proposed method performs well on low SNRs, especially at -10 dB. However as the SNR improves, the improvement over the baseline drops. This performance drop can be explained

by the reduced correlation between the squared error and the model uncertainty that is observed in Fig. 3.

Figure 3 plots the correlation between the predictive variance and the squared error (SE) of the estimated output frames for all the five MC models, for speech with white noise. The uncertainty is computed by taking the trace of the covariance matrix of each frame [25]. The plots show the weakening of the correlation between the SE and model uncertainty as the SNR improves. The correlation is strong for -10 and -5 dB and is weak for the values of SNR (0, 5 and 10 dB). This pattern of the correlation plots need further exploration. This matches with our results, since we find that there is not much improvement over the baseline model as the SNR increases. However, the values are still comparable to those of the single model scheme. This observation matches with that of [25], as the test data which are far from training set are likely to be more uncertain as the network is less adapted to them.

From the results in Table I and Fig. 3, the uncertainty based model selection shows promise of being potentially useful, especially in those cases, where the correlation between the model uncertainty and the square error is strong. Further analysis is needed to study the varying strength of this correlation. We would also like to learn the relationship between correlation and squared error better, so that the model can be selected in a risk minimization paradigm. Each model can be trained on a different group of noises and still this algorithm has the potential to be useful. Our experiments show that the proposed method gives a modest improvement in performance. The correlation plots show the potential use of the algorithm in real world noisy scenario, where the statistics of the training environment has a high mismatch with the application scenario. This needs to be explored further. The idea of MC dropout could be applied to any standard dropout network to explore the possibility of further improving the performance.

TABLE II
ENHANCEMENT PERFORMANCE OF A SINGLE DNN-MC DROPOUT MODEL (MC) ON *seen* NOISE AND *unseen* SNRS IN TERMS OF SUM SQUARED ERROR (SSE) AND SEGMENTAL SNR (SSNR).

		Factory2		
SNR (dB)	**Metric**	**Noisy input**	**Baseline**	**MC**
-10	**SSE x10^4**	4.13	0.0467	**0.0461**
	SSNR	-8.5	1.0	1.0
-5	**SSE x10^4**	1.29	0.0198	**0.0197**
	SSNR	-6.7	3.05	3.08
0	**SSE x10^3**	4.01	0.104	0.104
	SSNR	-4.1	5.1	5.1
5	**SSE x10^3**	1.24	0.069	0.069
	SSNR	-0.9	7.1	7.1
10	**SSE x10^2**	3.82	0.56	0.55
	SSNR	2.6	8.9	8.9

C. Performance impact

For MC dropout DNN models we experimented by adding dropout at various depths and found that using dropout in the final layer alone is effective [25]. Since dropout is added only in the final layer, the additional time required for drawing stochastic samples is marginal as the rest of the layers can be shared.

V. CONCLUSION AND FUTURE WORK

In this work, we propose two novel techniques that use dropout as a Bayesian estimator to improve the generalizability of DNN based speech enhancement algorithms. The first method uses the empirical mean of multiple stochastic passes through a DNN-MC dropout model trained on multiple noises to obtain the enhanced output. Our experiments show that this technique results in a better enhancement performance, especially on unseen noisy conditions. The second method looks at the potential application of the model uncertainty as an estimate of squared error (SE), for frame-wise selection of one out of multiple DNN models. While the experiments on validating this technique give only marginal improvement in some cases, the pattern of correlation between SE and model uncertainty, calls for further study. A particularly interesting line of study would include using complex functions that use the model uncertainty to arrive at the optimal model for each frame. This is the first study on the effectiveness of MC dropout for speech enhancement to the best of our knowledge. The main purpose of this work is to see the effectiveness of MC dropout over standard dropout models and hence could be implemented on any state of the art system employing dropout.

REFERENCES

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

[2] Y. Lu and P. C. Loizou, "A geometric approach to spectral subtraction," *Speech communication*, vol. 50, no. 6, pp. 453–466, 2008.

[3] V. Stahl, A. Fischer, and R. Bippus, "Quantile based noise estimation for spectral subtraction and Wiener filtering," in *Acoustics, Speech, and Signal Processing. ICASSP. Proceedings. Int. Conf.*, vol. 3. IEEE, 2000, pp. 1875–1878.

[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. 1109–1121, 1984.

[5] R. Martin, "Speech enhancement based on minimum mean-square error estimation and supergaussian priors," *IEEE transactions on speech and audio processing*, vol. 13, no. 5, pp. 845–856, 2005.

[6] J. S. Erkelens, R. C. Hendriks, R. Heusdens, and J. Jensen, "Minimum mean-square error estimation of discrete Fourier coefficients with generalized gamma priors," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 6, pp. 1741–1752, 2007.

[7] B. Yegnanarayana, C. Avendano, H. Hermansky, and P. S. Murthy, "Speech enhancement using linear prediction residual," *Speech communication*, vol. 28, no. 1, pp. 25–42, 1999.

[8] W. Jin and M. S. Scordilis, "Speech enhancement by residual domain constrained optimization," *Speech Communication*, vol. 48, no. 10, pp. 1349–1364, 2006.

[9] P. Krishnamoorthy and S. M. Prasanna, "Enhancement of noisy speech by temporal and spectral processing," *Speech Communication*, vol. 53, no. 2, pp. 154–174, 2011.

[10] S. Srinivasan, J. Samuelsson, and W. B. Kleijn, "Codebook driven short-term predictor parameter estimation for speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 1, pp. 163–176, 2006.

[11] Y. Ephraim, "A bayesian estimation approach for speech enhancement using hidden markov models," *IEEE Transactions on Signal Processing*, vol. 40, no. 4, pp. 725–735, 1992.

[12] H. Sameti, H. Sheikhzadeh, L. Deng, and R. L. Brennan, "HMM-based strategies for enhancement of speech signals embedded in nonstationary noise," *IEEE Transactions on Speech and Audio processing*, vol. 6, no. 5, pp. 445–455, 1998.

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

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

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

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

[17] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

[18] 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 *13th Annual Conf. , International Speech Communication Association*, 2012.

[19] 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.

[20] Y. Wang and D. Wang, "Towards scaling up classification-based speech separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 7, pp. 1381–1390, 2013.

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

[22] G. E. Dahl, T. N. Sainath, and G. E. Hinton, "Improving deep neural networks for LVCSR using rectified linear units and dropout," in *Acoustics, Speech and Signal Processing (ICASSP), International Conference on*. IEEE, 2013, pp. 8609–8613.

[23] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[24] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *International Conference on Machine Learning*, 2016, pp. 1050–1059.

[25] A. Kendall and R. Cipolla, "Modelling uncertainty in deep learning for camera relocalization," in *Robotics and Automation (ICRA), International Conference on*. IEEE, 2016, pp. 4762–4769.

[26] P. Papadopoulos, A. Tsiartas, and S. Narayanan, "Long-term snr estimation of speech signals in known and unknown channel conditions," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 12, pp. 2495–2506, 2016.

[27] Z.-Q. Wang, Y. Zhao, and D. Wang, "Phoneme-specific speech separation," in *IEEE Inter. Conf., Acoustics, Speech and Signal Processing*, 2016.

[28] P. M. Nazreen, A. G. Ramakrishnan, and P. K. Ghosh, "A class-specific speech enhancement for phoneme recognition: A dictionary learning approach," *Proc. Interspeech*, pp. 3728–3732, 2016.

[29] ——, "A joint enhancement-decoding formulation for noise robust phoneme recognition," *14th IEEE India Council Inter. Conf. (INDICON)*, 2017.

[30] 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, p. 27403, Feb. 1993.

[31] A. Varga and H. J. M. Steeneken, "Assessment for automatic speech recognition ii: Noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Commun.*, vol. 12, no. 3, pp. 247–251, Jul. 1993.

[32] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," *Proc. the ICLR 2015*, pp. 1–13, 2015.

[33] 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.