IMPROVED CEPSTRA MINIMUM-MEAN-SQUARE-ERROR NOISE REDUCTION ALGORITHM FOR ROBUST SPEECH RECOGNITION

Jinyu Li, Yan Huang, and Yifan Gong

Microsoft Corporation, One Microsoft Way, Redmond, WA 98052, U.S.A.

ABSTRACT

In the era of deep learning, although beam-forming multi-channel signal processing is still very helpful, it was reported that single-channel robust front-ends usually cannot benefit deep learning models because the layer-by-layer structure of deep learning models provides a feature extraction strategy that automatically derives powerful noise-resistant features from primitive raw data for senone classification. In this study, we show that the single-channel robust front-end is still very beneficial to deep learning modelling as long as it is well designed. We improve a robust front-end, cepstra minimum mean square error (CMMSE), by using more reliable voice activity detector, refined prior SNR estimation, better gain smoothing and two-stage processing. This new front-end, improved CMMSE (ICMMSE), is evaluated on the standard Aurora 2 and Chime 3 tasks, and a 3400 hour Microsoft Cortana digital assistant task using Gaussian mixture models, feed-forward deep neural networks, and long short-term memory recurrent neural networks, respectively. It is shown that ICMMSE is superior regardless of the underlying acoustic models and the scale of evaluation tasks, with 25.46% relative WER reduction on Aurora 2, up to 11.98% relative WER reduction on Chime 3, and up to 11.01% relative WER reduction on Cortana digital assistant task, respectively.

Index Terms— *cepstra minimum mean square error, noise robustness, deep neural networks*

1. INTRODUCTION

Environment robustness in automatic speech recognition (ASR) remains a difficult problem despite many years of research [1][2][3][4][5][6][7][8]. The deep learning based acoustic model technologies [9][10][11][12][13][14] bring new challenges to conventional noise-robustness technologies which are very well studied in the Gaussian mixture model (GMM) era. The robust front-end is a necessary component to maintain the accuracy of the traditional GMM-based ASR recognizer in noisy environments. However, for the deep learning based ASR systems, although beam-forming multi-channel signal processing is still very helpful [15][16][17][18], it was reported that single-channel robust front-ends cannot benefit deep learning models [19][20] in multi-style training setups. With the excellent modeling power of deep neural networks (DNNs), the DNN-based acoustic models can easily match state-of-the-art performance of GMM systems without any explicit noise compensation. This is because its layer-by-layer structure provides a feature extraction strategy that automatically derives powerful noise-resistant features from primitive raw data for senone classification, resulting in good noise-invariance property [8][21].

In this study, we show that the single-channel robust front-end is still very beneficial to deep learning models as long as it is well designed. Particularly, we work on improving a robust front-end named cepstra minimum mean square error (CMMSE) [22] which is very effective in dealing with noise when using the GMM-based acoustic models. However, it was shown that there was almost no improvement [19] from the standard front-end on a noise-robustness task Aurora 4 [23]. We elaborate how the components of CMMSE can be modified so that this front-end method can be redesigned into a powerful one. The improvement includes a better voice activity detection component which helps noise spectrum estimation, a refined prior SNR estimation for converged filter-bank gain, better gain smoothing method and two-stage processing to further clean the residual noise. We call the new method as improved CMMSE (ICMMSE) and evaluate it on three tasks. On the standard Aurora 2 task [24] with GMM acoustic models, ICMMSE gets 25.46% relative WER reduction. On the Chime 3 task [25] with feed forward DNN acoustic models, ICMMSE gets up to 11.98% relative WER reduction. On the product-scale Microsoft Cortana digital assistant task with long short-term memory (LSTM) recurrent neural networks (RNNs) acoustic models, the improvement is up to 11.01%. Hence, we establish that ICMMSE is superior regardless of the underlying models and evaluation tasks.

2. CEPSTRA MINIMUM MEAN SQUARE ERROR

The cepstra minimum mean square error (CMMSE) algorithm was proposed by Yu et al. [22]. It distinguishes from the MMSE enhancement in log spectral amplitude [26] in that it develops a suppression rule that applies to the outputs of the Mel filter-banks in the power spectrum domain and to Mel-frequency cepstrum coefficients (MFCC) directly. Thus, the enhancement is directly targeted to the features for ASR. The solution to CMMSE for each element of the dimension-wise MFCC is the conditional expectation

$$\hat{c}_x(t,k) = E\{c_x(t,k)|\boldsymbol{m}_y(t)\} = \sum_b a_{k,b}\, E\{\log m_x(t,b)\,|\boldsymbol{m}_y(t)\}, \quad (1)$$

where $c_x(t,k)$ is the k-th MFCC coefficient at frame t, b is the Mel filter-bank index, $a_{k,b}$ are the discrete cosine transform coefficients, $\boldsymbol{m}_x(t)$ and $\boldsymbol{m}_y(t)$ are the output of the Mel filter-bank in the power-spectrum domain for the clean and noisy speech, respectively.

Given the additive assumption for speech and noise signal, together with the weak independent assumption between Mel filter-banks, Eq. (1) can be simplified as

$$\hat{c}_x(t,k) \approx \sum_b a_{k,b}\, E\{\log m_x(t,b)|m_y(t,b)\}. \quad (2)$$

Then, the problem is reduced to finding the log-MMSE estimator of the Mel filter-bank's output

$$\hat{m}_x(t,b) \approx \exp\big(E\{\log m_x(t,b)\,|m_y(t,b)\}\big). \quad (3)$$

In the following, we briefly introduce the solution of CMMSE. Please refer [22] for details. Similar to a popular group of speech enhancement methods, CMMSE follows the 4-step processing:

- Voice activity detection (VAD): detects the speech probability at every time-frequency bin;

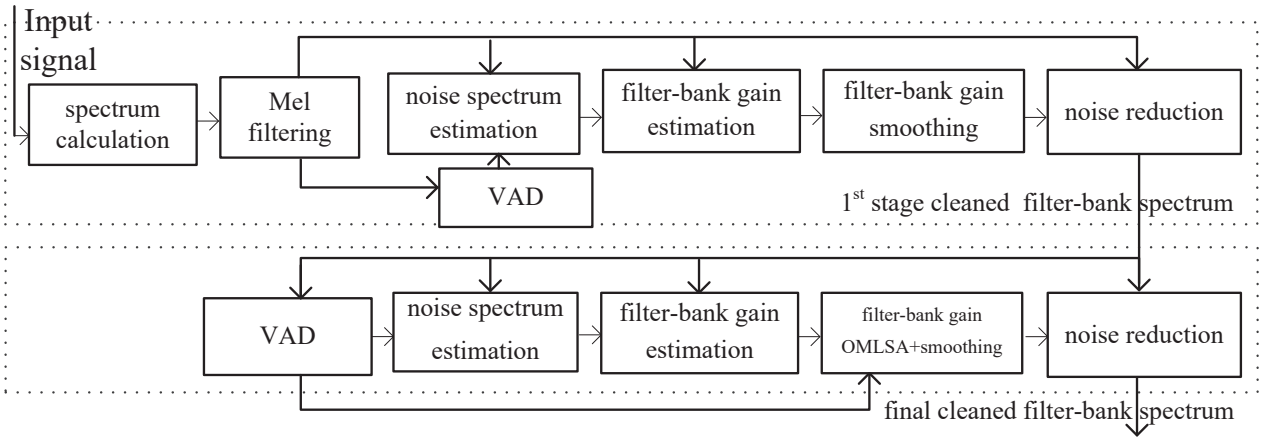

Figure 1. The flow chart of ICMMSE

- Noise spectrum estimation: use the estimated speech probability to update the estimation of noise spectrum;
- Gain estimation: use the noisy speech spectrum and the estimated noise spectrum to calculate the gain of every time-frequency bin;
- Noise reduction: apply the estimated gain to the noisy speech spectrum to generate the clean spectrum.

In the VAD part, CMMSE uses the method in [27] to detect the speech probability $p(t,b)$ in each filter-bank bin *b* and time *t*. Then, the noise power spectrum $m_n(t,b)$ is estimated using a minimum controlled recursive moving average (MCRA) noise tracker [27] as

$$m_n(t,b) = \alpha * m_n(t-1,b) + (1.0-\alpha) * m_x(t,b) \quad (4)$$

with

$$\alpha = \alpha_D + (1.0-\alpha_D) * p(t,b), \quad (5)$$

where $\alpha_D = 0.8$ in this study.

Following a similar approach in [26], the solution to Eq. (4) is

$$\widehat{m}_x(t,b) \approx G(t,b) m_y(t,b), \quad (6)$$

where $G(t,b)$ is the gain of time filter-bank bin with

$$G(t,b) = \frac{\hat{\xi}(t,b)}{1+\hat{\xi}(t,b)} \exp\left\{\frac{1}{2}\int_{v(t,b)}^{\infty} \frac{e^{-\tau}}{\tau} d\tau\right\}. \quad (7)$$

To calculate Eq. (7), the posterior SNR $\gamma(t,b)$ is first obtained as

$$\gamma(t,b) = \frac{m_y(t,b)}{m_n(t,b)} \quad (8)$$

and the prior SNR $\hat{\xi}(t,b)$is calculated using a decision-directed approach (DDA) [28]

$$\hat{\xi}(t,b) = \beta * G(t-1,b) * \gamma(t-1,b) + (1.0-\beta) * \xi(t,b) \quad (9)$$

where

$$\xi(t,b) = \max(\gamma(t,b) - 1, 0.0)$$

with $\beta = 0.9$ in this study. Then we have

$$v(t,b) = \frac{\hat{\xi}(t,b)}{1+\hat{\xi}(t,b)} \gamma(t,b). \quad (10)$$

Eq. (6) is used to estimate the cleaned power spectrum $\widehat{m}_x(t,b)$ of every time filter-bank bin. MFCCs can be obtained by plugging $\widehat{m}_x(t,b)$ into Eq. (2).

3. IMPROVED CEPSTRA MINIMUM MEAN SQUARE ERROR

In this section, we introduce how we develop the ICMMSE by improving the different components of CMMSE with more advanced algorithms. Specifically, an improved version of MCRA is used to get more accurate speech probability in each time filter-bank bin. Second, a refined prior SNR estimation benefits the gain estimation. Third, gain smoothing is employed to smooth the gains cross filter-banks. Lastly, two-stage processing is used to further clean the residual noise after the first stage processing. A high-level diagram of the ICMMSE is presented in Figure 1. We will present the details in the rest of this section.

CMMSE relies on MCRA [27] for speech probability estimation in VAD which is then used to estimate noise spectrum with Eq. (4). Reliable speech probability estimation is critical to the estimation of noise spectrum, which affects the accuracy of the posterior SNR estimation in Eq. (8). In [29], an improved MCRA (IMCRA) algorithm is shown to outperform MCRA. Therefore, we use IMCRA to estimate the speech probability $p(t,b)$ in each time filter-bank bin. Then, the noise power spectrum is still estimated with Eq. (4). The IMCRA process is much more complicated than MCRA. Due to limited space, we do not describe it in this paper. The readers can refer [29] for the detailed implementation.

The DDA prior estimation in Eq. (9) is a weighted sum of the SNRs in the previous and current frames. After estimating $G(t,b)$ in Eq. (7), we propose to do further gain estimation by using $G(t,b)$ to re-estimate prior SNR so that we can get a converged gain:

$$\dot{\xi}(t,b) = G(t,b) * \gamma(t,b)$$

$$\dot{v}(t,b) = \frac{\dot{\xi}(t,b)}{1+\dot{\xi}(t,b)} \gamma(t,b)$$

$$\dot{G}(t,b) = \frac{\dot{\xi}(t,b)}{1+\dot{\xi}(t,b)} \exp\left\{\frac{1}{2}\int_{\dot{v}(t,b)}^{\infty} \frac{e^{-\tau}}{\tau} d\tau\right\}$$

We call this process as refined prior SNR estimation.

Another improvement of ICMMSE is the estimation of gain $G(t,b)$ applied to the power spectrum of filter-bank in Eq. (6). In [30], an optimally-modified log-spectral amplitude (OMLSA) speech estimator is shown to be superior. To apply it to the time filter-bank bin, we modify the filter-bank gain with

$$\hat{G}(t,b) = \dot{G}(t,b)^{p(t,b)} G_0^{1-p(t,b)} \quad (11)$$

Here, G_0 is the minimum gain used to compress noise and is set as 0.1 in this study. There are two special cases when current speech probability $p(t,b)$ is 0 and 1.

$$\hat{G}(t,b) = \dot{G}(t,b) \quad if\ p(t,b) = 1$$

$$\hat{G}(t,b) = G_0 \quad if\ p(t,b) = 0$$

This means that if we are confident with the strong speech presence, we can use the estimate gain to clean noise. In contrast, if we are confident that current spectrum is only for noise, we can heavily clean it. For the intermediate cases, we do both operations.

A concern of OMLSA in Eq. (11) is that it relies on the speech probability which may not be accurately estimated when strong noise is present. Considering the fact that applying $\dot{G}(t,l)$ directly

to the noisy Mel filter-bank power spectrum may not be optimal because the number of Mel filter-banks is much less than the number of linear frequency bins, we propose a cross filter-bank gain smoothing method by averaging the predicted gain in neighborhood filter-banks as

$$\hat{G}(t,b) = (\dot{G}(t,b-1) + \dot{G}(t,b) + \dot{G}(t,b+1))/3 \quad (12)$$

Finally, the noise reduction process is not perfect due to the factors such as imperfect noise estimation. There is still residual noise in the cleaned spectrum. Inspired by the two-stage processing in ETSI advanced front-end [31], a second stage noise reduction can be used to further reduce the noise by using the cleaned speech spectrum in the first stage as the input to the second stage as shown in Figure 1. Most components in the second stage are the same as those in the first stage, except that gain smoothing in Eq. (12) is used in the first stage while we use OMLSA in Eq. (11) together with gain smoothing for the gain modification in the second stage because the residual noise has less impact to speech probability estimation after the first stage noise reduction. The speech probability obtained from the VAD module is only used for noise spectrum estimation in the first stage. In contrast, it is used in both the noise spectrum estimation and the OMLSA gain modification in the second stage. Note that all the processing in ICMMSE is on the power spectrum of time filter-bank bin which is the acoustic feature for ASR, while the traditional feature enhancement methods [26][27][28][29][30] work on the magnitude of linear frequency bin.

4. EXPERIMENTS

The effectiveness of the proposed ICMMSE algorithm is evaluated with three tasks. The first is a standard digit recognition task, Aurora 2 [24], with GMM modeling. We will show how ICMMSE can improve the CMMSE algorithm by breaking down the contribution from each ICMMSE component. The second task is the standard far-talk Chime 3 task [25] trained with feed-forward DNNs using tens of hours training data. The third task is a product-scale Microsoft Cortana digital assistant task using a long short-term memory (LSTM) [32][33] recurrent neural networks (RNNs) trained with 3400 hours of live speech data. We will show that ICMMSE is superior regardless of the underlying models and evaluation tasks.

4.1 Aurora 2

The multi-style training set which consists of 8440 multi-style training utterances is used to train the standard "simple backend" HMM model [24]. All digits are modeled with 16 states, with strict left-to-right structures. Each state is modeled by a GMM with 3 Gaussians. In addition, there is one "silence" model which consists of 3 states and each state is modeled by a GMM with 6 Gaussians.

Table 1: WER comparison of different front-ends on Aurora2. The models are trained with multi-style data.

	Raw	CMMSE	1-stage ICMMSE	2-stage ICMMSE
Clean	1.39	1.48	1.20	1.10
20db	2.69	2.21	1.99	1.85
15db	3.6	3.21	2.91	2.71
10db	6.04	5.65	5.30	4.97
5db	14.38	13.01	12.59	11.57
0db	43.41	38.36	34.02	32.06
-5db	75.93	72.8	68.47	67.45
Avg. (0-20 db)	13.67	12.17	10.87	10.19

The test material consists of three sets of distorted utterances. Set-A and set-B contain eight different types of additive noise while set-C contains two different types of noise and additional channel distortion. Each type of noise is added into a subset of clean speech utterances, with seven different levels of signal to noise ratios (SNRs). This generates seven subgroups of test sets for a specified noise type, with clean, 20db, 15db, 10db, 5db, 0db, and -5db SNRs. Following the standard evaluation of Aurora 2, we report average WER defined as the average of WERs of the 0-20db SNR test sets.

Table 1 shows the detailed results of baseline MFCC feature, and the MFCC features extracted with CMMSE and ICMMSE. CMMSE yields 10.97% relative WER reduction against the baseline MFCC feature. Even with 1-stage processing, ICMMSE is better than CMMSE in all conditions, and the 2-stage ICMMSE can reduce the average WER significantly from CMMSE, with 16.27% relative improvement from CMMSE, or 25.46% relative WER reduction from the baseline raw feature.

In Table 2, we show how CMMSE is evolved into ICMMSE by incrementally adding or replacing some components. The baseline CMMSE has 12.17% WER. After replacing the MCRA component with IMCRA, we get 4.19% relative WER reduction, showing the importance of voice active detection which affects the noise spectrum estimation. As mentioned in Section 3, one important difference between ICMMSE (or CMMSE) and the traditional method is that the gain is applied to the power spectrum in ICMMSE (or CMMSE) while the gain is applied to the magnitude spectrum in the traditional methods. Here, we also try to apply the gain to magnitude spectrum, and get 11.97% WER, about 2.66% relative increase from the 11.66% WER when applying the gain to power spectrum.

Using OMLSA estimation further improves the WER to 10.99% and with the refined prior we reach 10.87% WER. As discussed in Section 3, OMLSA may be too aggressive in the first stage processing. After replacing it with gain smoothing in Eq. (12), we obtain further accuracy improvement with 10.74% WER. Finally, we further clean the spectrum with the second stage ICMMSE processing and obtain 10.19% WER. In the following subsections, we will just use the 2-stage ICMMSE method as the default ICMMSE front-end given that we have shown that all the components in ICMMSE are helpful.

Table 2: Average WER (0-20db) comparison of different robust front-end methods, showing how CMMSE is evolved to ICMMSE step by step.

Method	Avg. WER
CMMSE	12.17
+ IMCRA	11.66
+OMLSA	10.99
+refined prior SNR	10.87
-OMLSA +gain smoothing (1-stage ICMMSE)	10.74
+2nd stage processing (2-stage ICMMSE)	10.19

4.2 Chime 3

The Chime 3 task is a scenario targeting the performance of ASR in a real-world, commercially-motivated scenario [13]. The data is recorded using a 6-channel microphone array mounted on a tablet.

The training data consists of 1600 real noisy utterances and 7138 simulated utterances. The real data is recorded in different live environments. The simulated data is obtained by mixing clean utterances into different background recordings. For both real and

simulated data, four environments have been selected: café (CAF), street junction (STR), public transport (BUS), and pedestrian area (PED). The development test set has 1640 real and 1640 simulated utterances. All utterances are selected from the WSJ0 corpus.

We train a fully connected deep neural network (DNN) on single-channel noisy far-talk speech. The DNN has 7-hidden layers, each with 2048 hidden units. The input consists of a 2640-dim feature vector formed by 80-dim log Mel-filter-bank (LFB) feature and its accelerating feature components with a context window of 11 frames (80*3*11=2640). The output layer has 3012 tied-triphone states (senones). We adopt the RBM pre-training [9] before the fine-tuning of the full network using the cross-entropy criterion.

The models are evaluated in single noisy channel and multi-channel enhanced speech provided by the Chime 3 speech challenge. Furthermore, we evaluate the different front-ends in clean testing condition. All test is conducted using the real development test sets. The language model (LM) used for decoding is the standard trigram from the WSJ corpus.

Table 3 summarizes the WER comparison of raw, CMMSE, and ICMMSE LFB features. In the single noisy channel or enhanced testing conditions, CMMSE and ICMMSE yield 8.60% and 11.72% or 7.42% and 11.98% WER reductions against the baseline raw LFB. Both CMMSE and ICMMSE outperform the raw LFB features in both far-field testing conditions. ICMMSE consistently yields additional accuracy gain against CMMSE. We further compare the performance on the clean testing condition. CMMSE has slight accuracy degradation against the raw LFB, nevertheless ICMMSE wins over CMMSE with small gain.

Table 3: WER comparison of raw, CMMSE, and ICMMSE on Chime3 task. The back-end is DNN trained from single noisy far-talk channel data. The models are evaluated on clean, noisy, and enhanced far-talk speech using real dev test sets. The results in the brackets are relative WER reductions from the raw feature.

	Raw	**CMMSE**		**ICMMSE**	
	WER	WER	*Relative Improve.*	WER	*Relative Improve.*
Clean	7.56	7.64	(-1.06)	7.41	(1.98)
Noisy	18.95	17.32	(8.60)	16.73	(11.72)
Enhanced	23.71	21.95	(7.42)	20.87	(11.98)

It is to be noted that we also experimented on the different front-ends using LSTM-RNN acoustic model with almost identical accuracy comparison results. Further experiments on maximum mutual information (MMI) sequence training result in similar performance comparison across different front-ends. The enhanced testing provided by the Chime3 challenge (MVDR based beamforming) is known to generate worse accuracy performance on the real test [13], as can be seen in the table above. Nevertheless, this does not affect our conclusion on the front-end comparison.

4.3 Cortana Task

We further evaluate the front-ends with a product scale Microsoft Cortana digital assistant task, trained with 3400hr live US English data. A trigram LM is used for decoding with around 8M ngrams. The test sets are from the same Microsoft Cortana task with three SNR conditions: 20db above, 10-20db, and 0-10db. The test sets in all conditions contain around 167k words, which guarantees the statistical significance of reported improvement.

The acoustic feature is the 80-dimensional static LFB. The baseline LFB features are extracted from the standard LFB generation without any robust front-end (i.e., raw LFB), and from CMMSE and ICMMSE robust front-ends. The training and testing use consistent features.

All acoustic models are 4-layer LSTM-RNNs with 5980 senones, and were trained to minimize the frame-level cross-entropy criterion. LSTM-RNNs have been shown to be superior than the feed-forward DNNs [33], which we previously verified with our Cortana task [34]. The LSTM-RNNs are modeled after the one described in [33] with the frame skipping strategy to reduce the runtime cost [34]. Each LSTM layer has 1024 hidden units and the output size of each LSTM layer is reduced to 512 using a linear projection layer. There is no frame stacking, and the output HMM state label is delayed by 5 frames as in [33]. When training LSTM, the backpropagation through time (BPTT) [35] step is 20.

Table 4 compares the WER of different front-ends in 20db above, 10-20db, and 0-10db testing conditions. It is interesting to see that in the 20db above condition, both ICMMSE and CMMSE outperform raw feature, with more than 3% relative WER reduction. The power of robust front-end is clear in the noisy conditions. CMMSE gets 4.66% and 3.81% relative WER reduction from the raw feature in the 10-20db and 0-10db conditions respectively, while ICMMSE gets 11.01% and 8.58% relative WER reduction.

It is interesting to see that compared to the case of using GMM acoustic models, the ICMMSE's improvement is halved when using the deep learning acoustic models. This indicates that the powerful layer-by-layer structure in deep learning models really make it very challenging for robust front-end to maintain improvement.

Table 4: WER comparison of different front-ends on Microsoft Cortana digital assistant task. The results in the brackets are relative WER reductions from the raw feature.

	Raw	**CMMSE**		**ICMMSE**	
	WER	WER	*Relative Improve.*	WER	*Relative Improve.*
20db above	13.17	12.64	(4.02)	12.71	(3.49)
10-20db	20.8	19.83	(4.66)	18.51	(11.01)
0-10db	27.03	26.00	(3.81)	24.71	(8.58)

5. CONCLUSIONS

In this paper, we proposed a new robust front-end called ICMMSE which improves the previous CMMSE front-end with several advanced components. The IMCRA algorithm helps to generate more accurate speech probability in each time filter-bank bin so that more reliable noise spectrum estimation can be obtained. The refined prior SNR estimation helps to get a converged gain. Either cross filter-bank gain smoothing or OMLSA is helpful to further modify the gain function. Finally, the two-stage processing helps to reduce the residual noise after the first-stage processing. It is shown in the experiment section that all these new components help to improve the ASR accuracy from the CMMSE front-end.

We compared different front-ends on the standard Aurora 2 task using Gaussian mixture models, the standard Chime 3 task using feed-forward DNNs, and a 3400 hour Microsoft Cortana digital assistant task using LSTM-RNNs, respectively. It is shown that ICMMSE is superior regardless of the underlying acoustic models and the scale of evaluation tasks, with 25.46% relative WER reduction on Aurora 2, up to 11.98% relative WER reduction on Chime 3, and up to 11.01% relative WER reduction on Cortana digital assistant task, respectively. This demonstrated that, while DNN models are more robust to noise, a well-designed robust front-end is still very helpful to deep learning acoustic models.

7. REFERENCES

[1] A. Acero, *Acoustical and Environmental Robustness in Automatic Speech Recognition*, Cambridge University Press, 1993.

[2] J.C. Junqua, and J.P. Haton, *Robustness in Automatic Speech Recognition: Fundamentals and Applications*. Kluwer Academic Publishers, 1995.

[3] Y. Gong, "Speech recognition in noisy environments: A survey," *Speech Communication*, vol. 16, pp. 261–291, 1995.

[4] J. Droppo and A. Acero, "*Environmental robustness*," in Handbook of Speech Processing, J. Benesty, M. M. Sondhi, and Y. Huang, Eds., chapter 33. Springer, 2008.

[5] M. J. F. Gales, "Model-based approaches to handling uncertainty," in *Robust Speech Recognition of Uncertain or Missing Data: Theory and Application*, pp. 101–125. Springer, 2011.

[6] T. Virtanen, R. Singh, and B. Raj eds. Techniques for noise robustness in automatic speech recognition, Wiley, 2012.

[7] J. Li, L. Deng, Y. Gong and R. Haeb-Umbach, "An Overview of Noise-Robust Automatic Speech Recognition," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 4, pp. 745-777, 2014.

[8] J. Li, L. Deng, R. Haeb-Umbach, and Y. Gong, *Robust Automatic Speech Recognition: A Bridge to Practical Applications*, Morgan Kaufmann Press, 2015.

[9] D. Yu, L. Deng, and G. Dahl, "Roles of pretraining and fine-tuning in context-dependent DBN-HMMs for real-world speech recognition," in *Proc. NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2010.

[10] T. N. Sainath, B. Kingsbury, B. Ramabhadran, P. Fousek, P. Novak, and A. Mohamed, "Making deep belief networks effective for large vocabulary continuous speech recognition," in *Proc. Workshop on Automatic Speech Recognition and Understanding*, pp. 30–35, 2011.

[11] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 20, no. 1, pp. 30–42, 2012.

[12] N. Jaitly, P. Nguyen, and V. Vanhoucke, "application of pretrained deep neural networks to large vocabulary speech recognition", in *Proc. Interspeech*, 2012.

[13] G. Hinton, L. Deng, D. Yu, G. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[14] L. Deng, J. Li, J. -T. Huang et al. "Recent advances in deep learning for speech research at Microsoft," in *Proc. ICASSP*, 2013.

[15] J. Du et al., "The USTC-iFlytek system for CHiME-4 challenge," CHiME-4 workshop, 2016.

[16] T. Menne et al., "The RWTH/UPB/FORTH system combination for the 4th CHiME challenge evaluation," CHiME-4 workshop, 2016.

[17] J. Heymann et al., "Wide residual BLSTM network with discriminative speaker adaptation for robust speech recognition," CHiME-4 workshop, 2016.

[18] H. Erdogan et al., "Multi-channel speech recognition: LSTMs all the way through," CHiME-4 workshop, 2016.

[19] M. L. Seltzer, D. Yu, and Y. Wang, "An investigation of deep neural networks for noise robust speech recognition," in *Proc. ICASSP*, pp. 7398–7402, 2013.

[20] B. Li, Y. Tsao, and K. C. Sim, "An investigation of spectral restoration algorithms for deep neural networks based noise robust speech recognition," in *Proc. Interspeech*, pp. 3002–3006, 2013.

[21] D. Yu, M. Seltzer, J. Li, J.-T. Huang, and F. Seide, "Feature Learning in Deep Neural Networks - Studies on Speech Recognition," in *International Conference on Learning Representations*, 2013.

[22] D. Yu, L. Deng, J. Droppo, J. Wu, Y. Gong, and A. Acero, "A minimum-mean-square-error noise reduction algorithm on Mel-frequency cepstra for robust speech recognition," in *Proc. ICASSP*, 2008.

[23] N. Parihar and J. Picone, "Aurora working group: DSR front end LVCSR evaluation AU/384/02," *Tech. Rep.*, Institute for Signal and Information Processing, Mississippi State Univ., 2002.

[24] H. G. Hirsch and D. Pearce, "The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Proc. ISCA ITRW ASR*, 2000.

[25] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, S, "The third CHiME speech separation and recognition challenge: Dataset, task and baselines," In *IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 504-511, 2015.

[26] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoustics, Speech and Signal Proc*, vol. ASSP-33, pp. 443–445, 1985.

[27] I. Cohen and B. Berdugo, "Noise estimation by minima controlled recursive averaging for robust speech enhancement," *IEEE signal processing letters*, *9*(1), pp.12-15, 2002.

[28] 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, pp1109-1121, 1984.

[29] I. Cohen, "Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging," *IEEE Trans. Speech and Audio Processing*, Vol. 11, No. 5, pp. 466-475, 2003.

[30] I. Cohen and B. Berdugo, "Speech enhancement for non-stationary noise environments," *Signal Processing*, Vol. 81, No. 11, pp. 2403-2418, 2001.

[31] ETSI, "Speech processing, transmission and quality aspects (STQ); distributed speech recognition; advanced front-end feature extraction algorithm; compression algorithms," 2002.

[32] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[33] H. Sak, A. Senior, F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling," in *Proc. Interspeech*, 2014.

[34] Y. Miao, J. Li, Y. Wang, S.X. Zhang, and Y Gong, "Simplifying long short-term memory acoustic models for fast training and decoding," in *Proc. ICASSP*, 2016.

[35] H. Jaeger, "Tutorial on training recurrent neural networks, covering BPPT, RTRL, EKF and the echo state network approach," *GMD Report 159*, GMD—German National Research Institute for Computer Science, 2002.