AN IDEAL HIDDEN-ACTIVATION MASK FOR DEEP NEURAL NETWORKS BASED NOISE-ROBUST SPEECH RECOGNITION

Bo Li, Khe Chai Sim

National University of Singapore
School of Computing, 13 Computing Drive, Singapore 117417

ABSTRACT

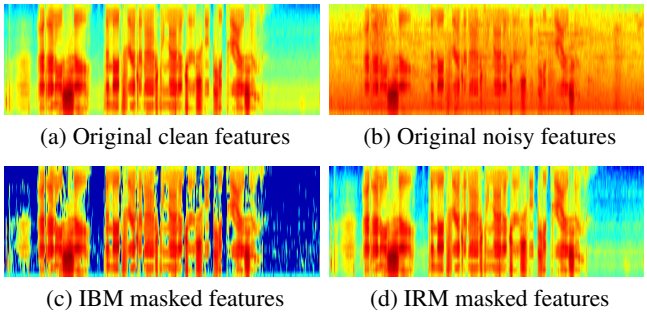
Deep neural networks (DNNs) are capable of modeling large acoustic variations. However, the performance on noisy data is still below humans' expectations. In this work, we present an ideal hidden-activation masking (IHM) approach to improve their noise robustness. This IHM is inspired by the existing spectral masking techniques. Instead of masking away the noise-dominant components in the spectral domain, we propose to discard DNNs' inconsistent hidden activations. The IHM is computed from the parallel data to identify hidden units that are immune to environment noise. DNNs then utilize it to improve their prediction robustness with the noise-invariant activations. Experimental results on the Aurora4 task have shown that the proposed IHM is both effective in reducing noise variations and robust to mask estimation errors.

Index Terms— Deep Neural Networks, Noise Robustness

1. INTRODUCTION

With the fast adoption of speech-based services, noise robustness of automatic speech recognition (ASR) systems is becoming more and more crucial to better user experiences in real world applications. Deep neural networks (DNNs) have shown a much better generalization capability than conventional Gaussian mixture models (GMMs) [1]. However, the performance of DNNs on speech from noisy environments is still far from humans' expectations. Addressing DNNs' noise robustness is attracting much interest.

In this work, we mainly focus on the speech-independent noise. To compensate the decreased intelligibility caused by noise, we could either enhance the target speech or reduce the interfering noise. In [2, 3], different feature enhancement techniques are borrowed from GMMs for DNNs; however, none of them could outperform the multi-style DNN baseline. [4, 5] adopt neural network models to directly reconstruct clean speech features from the noisy ones. They have no explicit noise assumption and are hence more dependent upon the training data to provide a reasonable sample of potential noise environments. Besides these feature denoising approaches, there are also some attempts to incorporate noise statistics into DNN models. In [6], a factorial hidden restricted Boltzmann machine is developed to explicitly model the noise distribution and how the noise affects the speech. However, due to the unobserved noise parameters, the inference is intractable and scaling exponentially with the number of hidden units. [7] treats the global mean and variance normalization (MVN) process as a single Gaussian generative front-end for DNNs and applies VTS to compensate it using noise estimation from the target test utterance. Due to the over simplicity of the single Gaussian-based compensation, it cannot outperform the per-utterance MVN baseline. Only when adaptive training is used does it yield moderate gains. In [8], the acoustic features are concatenated with noise parameters to train a "noise aware" DNN. However the grain is negligible and only after adopting the dropout fine-tuning [9] does it yield better performance.



(a) Original clean features (b) Original noisy features

(c) IBM masked features (d) IRM masked features

Fig. 1. Different spectrograms for the utterance "`440c0201`".

Recently, the spectral masking approach has shown some promising results in improving DNNs' noise robustness [10]. Motivated from humans' separation-prior-to-recognition speech perception process [11, 12], masks are adopted to separate speech from noise. An ideal binary mask (IBM) [13, 14] is used to identify each unit in a time-frequency (T-F) representation of the noisy signal as speech dominant or noise dominant. When applied in a direct masking manner [15, 16] , the IBM is used as a binary gain function to attenuate the energy within the noise-dominant T-F units. In [17], an ideal ratio mask (IRM) is developed and has been shown to outperform the IBM. In this work, we further extend the idea of masking into DNNs' hidden layers, which is referred to as the ideal hidden-activation mask (IHM). Instead of masking away the noise-dominant units, we discard the hidden units that generate inconsistent activations for speech from different noise conditions. The rest of the paper is organized as follows. In Section 2, the existing spectral masking techniques, *i.e.* the IBM and the IRM, are discussed. Section 3 details the proposed IHM. Evaluation results are presented in Section 4 and we conclude the paper in Section 5.

2. SPECTRAL MASKING

One straightforward explanation to DNNs' degradations on noisy data is the mismatch between the clean and noisy speech, which is illustrated in Fig. 1(a) and Fig. 1(b). To address this mismatch problem, the spectral masking technique aims to reduce the corruption noise in the power spectrum domain such that the extracted spectral/cepstral features are consistent across different noise environments. Two types of spectral masks are commonly used, namely the ideal binary mask and the ideal ratio mask.

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

2.1. Ideal Binary Mask (IBM)

The IBM [13, 14] is defined as:

$$m_{t,f}^{(\texttt{IBM})} = \begin{cases} 1 & \text{if } r_{t,f} > LC \\ 0 & \text{otherwise,} \end{cases} \quad \text{and } r_{t,f} = 10\log_{10}\frac{x(t,f)}{n(t,f)}, \quad (1)$$

where $r_{t,f}$ is the instantaneous local signal-to-noise ratio (SNR) of the T-F unit at the time frame t and the frequency channel f. $x(t,f)$ and $n(t,f)$ are the corresponding speech and noise energies. LC is a local SNR criterion and is set to -6dB as suggested in [18, 19].

2.2. Ideal Ratio Mask (IRM)

The IRM [17, 20] is defined as:

$$m_{t,f}^{(\texttt{IRM})} = (1 + \exp(-\gamma * (r_{t,f} - \beta)))^{-1} \quad (2)$$

where γ controls the slope of the sigmoid function and β corresponds to the LC. By tuning γ and β, we can control the range of SNRs to focus on while training the mask estimators. As suggested in [17], $\gamma = 0.2$ and $\beta = -6$ dB are adopted.

A visual comparison between the masked features (Fig. 1(c) and Fig. 1(d)) and the original clean and noisy features (Fig. 1(a) and Fig. 1(b)) could suggest that the masked features looks more similar to the clean speech. Moreover, the IRM seems capable of retaining more detailed information than the IBM.

3. IDEAL HIDDEN-ACTIVATION MASKING (IHM)

Speech data arises from the rich interaction of many sources. These factors interact in a complex way that complicates the recognition task. If we could identify and separate out these factors, we would largely ease the learning problem. The powerfulness of DNNs over GMMs in modeling large acoustic variations also comes from DNNs' high-level abstraction capabilities in identifying the underlying factors. With the guidance of the task specific supervisions at the final output layer, the distributed hidden representations at each layer try to encode only the underlying speech-dependent factors and discard noise factors in its input features. Using many layers' nonlinear transformations, DNNs could encode a rather complex relationship between the original acoustic features and the target classification labels. However, due to the commonly adopted gradient-based learning, the supervision strength decreases through many layers' back-propagation and the confusion increases. The layers near inputs hence have to maintain more redundancies to avoid missing any potential clues. When the testing data is similar to the training data, these redundant feature detectors have similar active levels as those have been seen during training and hence will not cause any problem. But when there are noise variations, they may become unexpectedly active and lead to probable performance degradations. We thus propose to mask away those unreliable feature detectors in DNN layers for improved noise robustness.

Due to the lack of the intuitive relationships between the hidden units and the target classification labels, we use the parallel speech data to guide the learning of noise-invariant hidden detectors. By comparing the hidden activations generated from the noisy and the corresponding clean speech, we could identify activations that are consistent between them; the corresponding feature detectors (*i.e.* the hidden units) will then be marked as noise-invariant. This mask is named the ideal hidden-activation mask (IHM). By applying this IHM, the hidden representations will become less noise-prone and the following DNN layers may easily yield correct predictions. The mathematical formulation of the IHM is defined as follows:

$$m_{l,t,f}^{(\texttt{IHM})} = \begin{cases} 1 & \text{if } s_{l,t,f} > \theta \\ 0 & \text{otherwise,} \end{cases} \quad (3)$$

$$s_{l,t,f} = \exp\{-\alpha * (h_{l,t,f}^{(\texttt{clean})} - h_{l,t,f}^{(\texttt{noisy})})^2\} \quad (4)$$

where $s_{l,t,f}$ denotes the similarity between the DNN's clean hidden activation, $h_{l,t,f}^{(\texttt{clean})}$, and the DNN's noisy hidden activation, $h_{l,t,f}^{(\texttt{noisy})}$, of the fth hidden unit in the lth hidden layer at the tth time frame. The two parameters α and θ controls the shape of the similarity curve and the threshold to decide whether a detector is noise-invariant. By default, we use the setting of $\alpha = 1.0$ and $\theta = 0.5$. The similarity curves (Equ. (4)) with different α values are plotted in Fig. 2. In practice, the lack of parallel data for testing requires the estimation of IHMs. In this work, a DNN-based mask estimator is learned with the training IHMs as supervision targets. During testing, we directly use $m_{l,t,f}^{(\texttt{IHM})} = s_{l,t,f}$ to alleviate potential errors in the mask estimation.

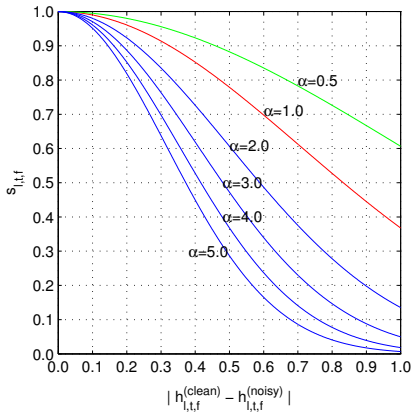


Fig. 2. The similarity function for the IHM.

The proposed IHM may look similar to the dropout technique [9], but they are different. The masking noise found in the dropout encourages a faster symmetry breaking by randomly discarding some of the hidden activations. However, the IHM is a deterministic way of identifying the noise-invariant hidden detectors.

4. EXPERIMENTS

In this section we present our investigation of the proposed IHM on the Aurora4 task [21]. The multi-style training data is used for training and the clean data is used only for computations of ideal masks. The complete test data consists of 14 subsets. Set 01 is clean and sets 02 $\sim$ 07 each is corrupted by one of the six different noise types (street traffic, train station, car, babble, restaurant, airport) at 5-15 dB SNR. Set 08 is filtered to incur channel distortions and sets 09 $\sim$ 14 further add one of the six noise types similarly. The noise is common across training and testing but the SNR differs, which is 10-20 dB for training. The dev data also consists of 14 subsets corresponding to the test data and is used for DNN's fine-tuning and IHM's parameter tunning.

4.1. Baseline

A context-dependent GMM-HMM system with 3257 senones is trained in a maximum likelihood manner. The per-utterance cepstral MVN normalized 39D MFCC features are used. This model is used to generate per-frame senone labels for the DNN training. A

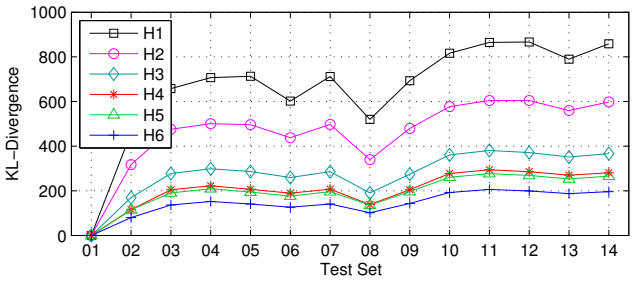


Fig. 3. The average test KL divergence between the noisy and clean hidden activations at different hidden layers of the baseline DNN.

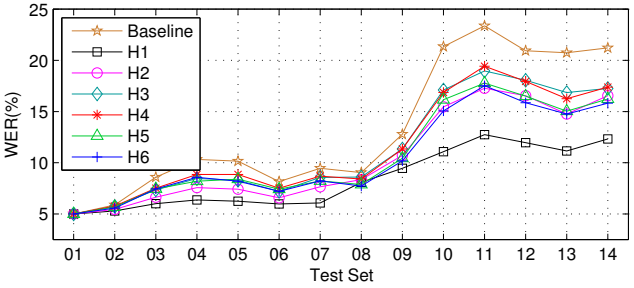


Fig. 4. The test WER performance of applying IHM ($\alpha = 1.0$ and $\theta = 0.5$) at different hidden layers of the baseline DNN.

6-hidden layer DNN is trained with an unsupervised pre-training followed by a supervised fine-tuning. The inputs are 11 adjacent frames of per-utterance MVN normalized 72D FBank features including the static, Δ and $\Delta\Delta$ statistics. We use 2048 hidden units at each hidden layer. The softmax output layer has 3257 units corresponding to the senones in the GMM-HMM system. All the decodings are performed with the WSJ0 bigram language model. This baseline DNN's performance is tabulated in the first row of Table 1. The performance degradation on noisy data is clearly observable by comparing the set 02 $\sim$ 07 to the set 01. With additional channel distortions, the error rate further increases by 2 to 3 times.

4.2. Ideal Hidden-Activation Masking (IHM)

One of the assumptions to adopt the IHM is that the DNN-generated hidden representations are not invariant to noise. Different noise in speech signals may cause variations in the hidden feature representations. To validate this assumption, we compare different hidden activations of the baseline DNN between each noisy set (02 $\sim$ 14) and the clean one (01) using KL-divergence in Fig. 3. It shows that the difference decreases dramatically while going deeper into the DNN. On one hand, this reflects the importance of adopting many layers for DNNs to obtain invariant feature representations; on the other hand, there are still variations caused by noise in the high level representations learned by DNNs and the lower layers are more prone to noise. To further improve DNNs' noise robustness, reducing the variations caused by noise in the hidden representations may be helpful.

Another assumption of the IHM is the redundancies in the hidden representations. By masking away some of the activations, there should be sufficient information left for classification. To justify this, we apply the proposed IHM with $\alpha = 1.0$ and $\theta = 0.5$ to different hidden layers of the baseline DNN. From the results in Fig. 4, all the IHMs have lower WERs on noisy test sets. Although there is not a clear trend consistent with the KL-divergence (Fig. 3), applying the IHM at the layer with the largest mismatch, *i.e.* H1, achieves the best performance. In the following study, we will focus only on the H1.

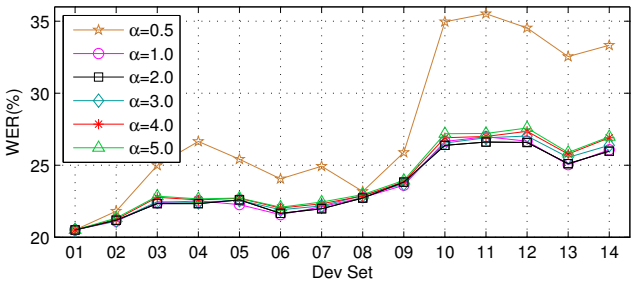


Fig. 5. The dev WER performance of applying IHM at H1 of the baseline DNN with different α values and fixed $\theta = 0.5$.

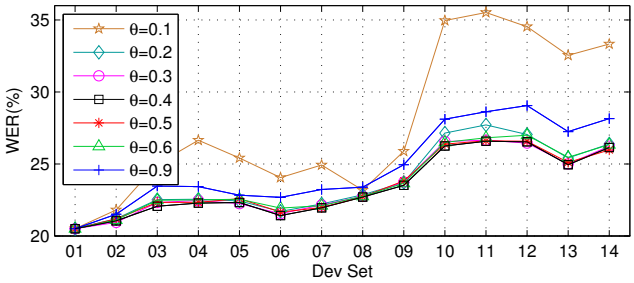


Fig. 6. The dev WER performance of applying IHM at H1 of the baseline DNN with different θ values and fixed $\alpha = 2.0$.

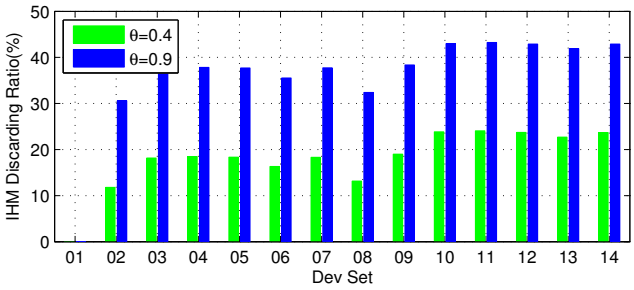


Fig. 7. The dev discarding ratios of active hidden features (> 0.001) by applying the IHM at H1 of baseline DNN with $\alpha = 2.0$.

Next, we tune the IHM parameters, α and θ, on the dev data. The results of varying α while fixing $\theta = 0.5$ are presented in Fig. 5. The differences are relatively small except for $\alpha = 0.5$ which actually uses no masks. That is because from Fig. 2, the configuration $\alpha = 0.5$ and $\theta = 0.5$ simply generates an all-one mask. Among all the values we investigate, $\alpha = 2.0$ is slightly better; we hence use this value for the following experiments. Next we vary the threshold parameter θ while fixing $\alpha = 2.0$. When $\theta = 0.1$, nothing will be masked away (Fig. 2). When $\theta = 0.9$, we mask away on average 35.8% of the active H1 hidden activations (those with values above 0.001) and we could still obtain an average 10.6% relative WER reduction. While changing the θ from 0.2 to 0.6 (Fig. 6), the WER performance has only small variations and reaches the minimum at $\theta = 0.4$, which has the average discarding ratio of 18.0%. The per dev set IHM activation discarding ratios for $\theta = 0.9$ and $\theta = 0.4$ are also compared in Fig. 7. Finally, by applying IHMs with $\alpha = 2.0$ and $\theta = 0.4$ on the test data, we can obtain an average WER of 8.2% and an average hidden activation discarding ratio of 18.9%.

4.3. Comparisons with IBM and IRM

In this set of experiments we compare our proposed IHM with the existing spectral masking techniques, namely the IBM and the IRM.

Table 1. Aurora4 WER (%) performance on each test set (01 $\sim$ 14) using different masks with different experiment configurations.

System		Test Set														Avg.
		01	02	03	04	05	06	07	08	09	10	11	12	13	14	
Baseline		5.0	5.9	8.6	10.3	10.2	8.2	9.5	9.0	12.8	21.4	23.4	20.9	20.7	21.2	13.4
E1	IBM	5.1	13.5	19.5	22.6	21.3	17.4	20.3	19.1	22.5	28.2	27.2	28.0	27.8	28.9	21.5
	IRM	5.1	5.6	5.9	5.8	5.7	5.7	5.6	7.4	9.2	8.3	7.1	8.2	8.3	8.7	**6.9**
	IHM	5.0	5.2	5.9	6.2	6.2	5.9	6.0	8.0	9.3	11.2	12.0	11.9	10.6	12.0	8.2
E2	IBM	4.8	5.3	5.8	6.0	6.4	5.8	6.4	9.4	8.4	8.4	8.0	8.4	8.0	8.8	7.1
	IRM	4.4	4.5	4.7	4.4	4.6	4.5	4.5	6.4	6.7	6.1	5.5	5.5	6.3	5.9	**5.3**
	IHM	5.0	4.9	5.5	5.8	5.8	5.3	6.2	6.9	6.9	9.1	10.5	9.8	9.4	10.2	7.2
E3	IBM	4.9	7.5	12.5	15.7	14.0	12.4	14.2	10.3	15.9	29.3	31.7	29.8	28.3	29.3	18.3
	IRM	4.5	6.2	10.3	12.3	11.4	9.9	11.2	9.3	15.7	25.9	28.1	26.3	25.3	26.4	15.9
	IHM	5.1	6.0	9.7	10.3	10.5	8.7	10.5	8.9	13.4	21.2	23.4	21.8	20.8	22.1	**13.7**
E4	IBM	4.6	5.7	9.2	11.0	10.7	9.3	9.8	8.4	12.7	23.7	25.0	22.6	22.0	22.8	14.1
	IRM	4.7	5.5	9.0	10.1	10.2	9.1	10.1	8.2	13.0	22.7	24.6	22.5	22.3	22.3	13.9
	IHM	4.9	5.8	8.8	10.0	9.9	8.4	9.7	8.8	12.1	20.8	23.4	21.1	20.0	20.7	**13.2**
E5	IBM	4.7	5.6	8.1	9.5	9.4	7.9	9.1	8.2	11.5	20.3	22.6	20.0	19.8	19.6	**12.6**
	IRM	4.9	5.6	8.3	9.8	9.3	7.8	9.4	8.0	12.0	20.6	22.7	20.2	19.7	19.5	12.7
	IHM	4.9	5.7	8.4	10.1	9.6	8.0	9.3	8.8	12.1	20.7	23.0	20.8	19.8	20.6	13.0

Firstly, the ideal masks are applied to the test sets only and evaluated with the baseline DNN. This is denoted as "E1" in Table 1. The real-valued IRM yields lower WERs than both the two binary masks due to the richness of its scaling based masking (Fig. 1(d)). Our IHM largely outperforms the IBM, of which the large degradation comes from the mismatch between the masked features and the training data (Fig. 1(c)). To reduce this mismatch we retrain the baseline DNN with ideally masked training data, which is denoted as "E2" in Table 1. With retraining, all the three masks have achieved further WER reductions. The IRM still performs the best and our IHM and the IBM have similar WERs. The dramatic change in IBM performance from "E1" to "E2" further confirms the differences between the masked and the original features.

The investigations of ideal masks could tell us the potential of different masks in removing noise corruptions. However, for real applications, the lack of ideal masks poses a great challenge for all the masking techniques. The errors in the estimated masks may even over-weigh the gains obtained. Finding the mask that is both effective in variation removal and robust to estimation errors is crucial to practical applications. We hence build three 6-hidden layer DNN-based mask estimators respectively. Details about the learning of the mask estimators could be found in [10]. In the "E3" part of Table 1, we first evaluate the estimated masks with the ideal masked DNN, *i.e.* the DNN used in "E2". All the masks degrade the performance and our proposed IHM has the minimum degradation. It suggests that the errors in the estimated masks are crucial. To address the mismatch between the ideal masks and the estimated masks, we retrain the baseline DNN with the estimated masks instead of the ideal ones. From the "E4" results of Table 1, our IHM performs the best and is the only one that improves the baseline DNN's performance. The improvement is also statistically significant at the level of $p = 0.05$ using the matched pair sentence segment word error method. By further comparing the relative WER reductions of these masks in Fig. 8, the spectral masking is preferable when the noise is simple (such as the car noise in set 02); but it degrades largely when the additive-noise assumption fails. The proposed IHM aims to identify the noise-invariant feature detectors and is hence more reliable across different noise types. On some sets (such as 03, 06, 07, 11 and 12), degradations have been observed for all the masks, which may require better mask estimations.

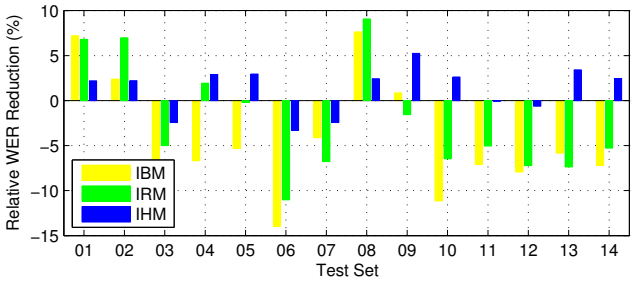


Fig. 8. The comparisons of the estimated IBM, IRM and IHM using relative test WER reductions from the baseline system.

In [10], although the estimated spectral masks cannot improve the baseline DNN's performance, they do provide complementary information to yield gains by averaging the two sets of posteriors. We thus average the posteriors generated from "E4" and the baseline respectively. The results reported in "E5" of Table 1 reconfirm the finding in [10] and further gains are achieved for our IHM.

5. CONCLUSIONS

In this paper, we propose the use of an ideal hidden-activation mask (IHM) at the first hidden layer of DNN acoustic models to further improve their noise robustness. Unlike the traditional spectral masking techniques such as the ideal binary masking (IBM) and the ideal ratio masking (IRM), the IHM operates at the DNN's distributed hidden representation space rather than the power spectral feature domain. The IBM and IRM aim to reduce the noise corruption and have the assumption of additive noise; while the IHM targets to discard noise-prone hidden units and has no noise type assumptions. Experimental results on the Aurora4 task have shown that the IHM is both effective in noise reduction and robust to mask estimation errors.

6. ACKNOWLEDGMENTS

This research is supported by the Singapore National Research Foundation under its International Research Centre @Singapore Funding Initiative and administered by the IDM Programme Office.

7. REFERENCES

[1] Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al., "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *Signal Processing Magazine, IEEE*, vol. 29, no. 6, pp. 82–97, 2012.

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

[3] 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*, 2013.

[4] O. Vinyals, S.V. Ravuri, and D. Povey, "Revisiting recurrent neural networks for robust ASR," in *Proc. ICASSP*, 2012.

[5] A.L. Maas, Q.V. Le, et al., "Recurrent neural networks for noise reduction in robust ASR," in *Proc. Interspeech*, 2012.

[6] S.J. Rennie, P. Fousek, and P.L. Dognin, "Factorial hidden restricted boltzmann machines for noise robust speech recognition," in *Proc. ICASSP*. IEEE, 2012.

[7] B. Li and K.C. Sim, "Noise adaptive front-end normalization based on vector taylor series for deep neural networks in robust speech recognition," in *Proc. ICASSP*, 2013.

[8] D. Yu, M.L. Seltzer, J.Y. Li, J.T. Huang, and F. Seide, "Feature learning in deep neural networks - a study on speech recognition tasks," in *Proc. ICLR*, 2013.

[9] G.E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R.R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," *arXiv:1207.0580*, 2012.

[10] B. Li and K.C. Sim, "Improving robustness of deep neural networks via spectral masking for automatic speech recognition," in *Proc. ASRU*. IEEE, 2013.

[11] J. Boldt, *Binary Masking & Speech Intelligibility*, Ph.D. thesis, Aalborg Universitet, 2011.

[12] D.L. Wang, G.J. Brown, et al., *Computational auditory scene analysis: Principles, algorithms, and applications*, 2006.

[13] D.L. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," *Speech separation by humans and machines*, 2005.

[14] U. Kjems, J.B. Boldt, et al., "Role of mask pattern in intelligibility of ideal binary-masked noisy speech," *The Journal of the Acoustical Society of America*, 2009.

[15] W. Hartmann, A. Narayanan, et al., "Nothing doing: Re-evaluating missing feature ASR," *Reconstruction*, 2011.

[16] W. Hartmann, A. Narayanan, E. Fosler-Lussier, and D.L. Wang, "A direct masking approach to robust asr," *Audio, Speech, and Language Processing, IEEE Transactions on*, 2013.

[17] A. Narayanan and D.L. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," in *Proc. ICASSP*. IEEE, 2013.

[18] D.S. Brungart, P.S. Chang, B.D. Simpson, and D.L. Wang, "Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation," *The Journal of the Acoustical Society of America*, vol. 120, pp. 4007, 2006.

[19] D.L. Wang, U. Kjems, M.S. Pedersen, J.B. Boldt, and T. Lunner, "Speech intelligibility in background noise with ideal binary time-frequency masking," *The Journal of the Acoustical Society of America*, vol. 125, pp. 2336, 2009.

[20] A. Narayanan and D.L. Wang, "Investigation of speech separation as a front-end for noise robust speech recognition," *OSU-CISRC-6/13-TR14*, 2013.

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