IMPROVING ROBUSTNESS OF DEEP NEURAL NETWORKS VIA SPECTRAL MASKING FOR AUTOMATIC SPEECH RECOGNITION

Bo Li, Khe Chai Sim

School of Computing
National University of Singapore
13 Computing Drive, Singapore 117417
li-bo@outlook.com, simkc@comp.nus.edu.sg

ABSTRACT

The performance of human listeners degrades rather slowly compared to machines in noisy environments. This has been attributed to the ability of performing auditory scene analysis which separates the speech prior to recognition. In this work, we investigate two mask estimation approaches, namely the state dependent and the deep neural network (DNN) based estimations, to separate speech from noises for improving DNN acoustic models' noise robustness. The second approach has been experimentally shown to outperform the first one. Due to the stereo data based training and ill-defined masks for speech with channel distortions, both methods do not generalize well to unseen conditions and fail to beat the performance of the multi-style trained baseline system. However, the model trained on masked features demonstrates strong complementariness to the baseline model. The simple average of the two system's posteriors yields word error rates of 4.4% on Aurora2 and 12.3% on Aurora4.

Index Terms— Deep Neural Network, Spectral Masking, Noise Robustness

1. INTRODUCTION

With the wide 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). However, the performance of DNNs on speech from unseen noise environments is still far from human expectations. Exploring the noise robustness of DNNs is attracting more interest.

Speech and noise are believed to be independent to each other. To compensate the decreased intelligibility caused by noise, we could either enhance the target speech or reduce the interfering noise. In the literature, various feature enhancement techniques aiming at improving speech intelligibilities have been developed. They are also one of the early attempts for DNNs due to their direct applicability [1, 2]. However, the performance improvement has only been seen for DNNs trained on clean signals. With multi-style data, slight degradations have been observed when using enhanced features. This may be attributed to the imperfect enhancement process that may discard useful speech information and meanwhile, bring in unwanted distortions. Besides the traditional enhancement algorithms, neural network models have also been trained on stereo data to directly reconstruct clean speech features, such as recurrent neural networks (RNNs) [3, 4]. They have shown improvements in matched noise conditions but also failed for unseen noises.

In human speech perception process, the human auditory system is believed to be capable of efficiently identifying and separating speech and noise prior to recognition [5]. Similarly, methods that using only the separated speech dominated time-frequency (T-F) units have seen many applications in robust ASR recently. The separation is usually done through binary masking. With stereo data, ideal binary masks (IBMs) [6] have been shown to largely improve intelligibilities of speech with background noises [7]. To suppress noises for ASR, IBMs are commonly used either by direct masking [8] or by performing reconstruction [9]. In direct masking, the noise dominate T-F units are discarded by the binary selection of IBMs; while in reconstruction, the speech energy for those units are estimated using information of the speech dominate units. The performance of both these two methods depends largely on the quality of IBM estimations. Various classification based algorithms for IBM prediction have thus been developed [10–13].

In this study, we investigate the effectiveness of spectral domain masking in the hybrid DNN-Hidden Markov Model (HMM) based ASR systems. Conventionally, an ensemble of human engineered features is required for IBM predictions [10]. While in [14], the authors find that features extracted automatically using a Gaussian-Bernoulli Restricted Boltzmann Machine (GRBM) perform even better. Similar trends have been observed in the acoustic modeling research. The log Mel filterbank (FBank) features [15] or the waveform signals [16] have been shown to outperform the human engineered Mel frequency cepstral coefficient (MFCC) features as inputs for DNNs. We hence justify the potential of estimating IBMs directly with FBank features. Two approaches, namely the state dependent and the DNN based IBM estimations, are proposed and the direct masking is adopted to filter the spectral features with estimated masks. To avoid possible errors brought by the mask binarization process, outputs of IBM estimators are directly used as soft masks, representing the expected probability for each T-F unit being speech dominated. Experiments on Aurora2 [17] and Aurora4 [18] noisy speech recognition tasks have shown that the estimated masks improve clean trained system but degrade the multi-style trained baseline, which has also been observed in [19]. However, the posteriors generated from it demonstrate strong complementariness to the baseline. A simple posterior interpolation yields much better performance than both of the two systems. The word error rates (WERs) of 4.4% on Aurora2 and 12.3% on Aurora4 achieved are both among the best performances of these datasets. The rest of the paper is organized as follows. The general spectral masking process is described in Section 2 and the proposed IBM estimation methods are detailed in Section 3. We present the experimental results in Section 4 and conclude the paper in Section 5.

2. SPECTRAL MASKING

The proposed integration of spectral masking in the hybrid DNN-HMM system is depicted in Figure 1. The masking process (light green shaded box in Figure 1) is performed on the power spectral features, which is simply an element-wise multiplication between the spectrum and the mask. The success of this system largely depends on the performance of the mask estimator.

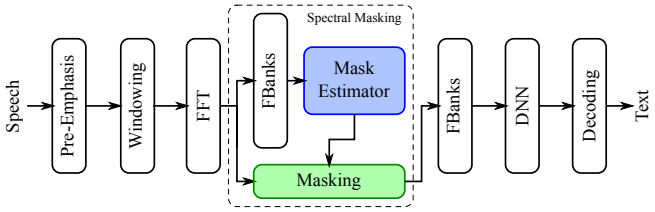
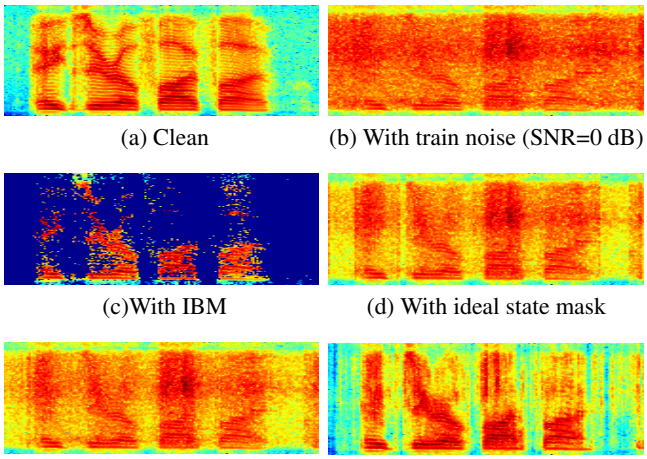


Fig. 1. The proposed system architecture for integrating spectral masking in the hybrid DNN-HMM system.

With stereo data, the IBM that labels each T-F unit of the noisy signal as speech dominant or noise dominant is computed using:

$$IBM(m,c) = \begin{cases} 1 & \text{if } SNR(m,c) > LC \\ 0 & \text{otherwise,} \end{cases} \quad (1)$$

where $SNR(m,c)$ represents the local signal-to-noise ratio (SNR) at time frame m and frequency channel c, and LC is a local SNR threshold. An example of IBM filtered noisy spectrogram is illustrated in Figure 2(c). Together with the unwanted noise T-F units, many speech details are also filtered away after masking by comparing it with the clean spectrum in Figure 2(a). This is why a feature reconstruction step is commonly adopted after masking.



(a) Clean (b) With train noise (SNR=0 dB)

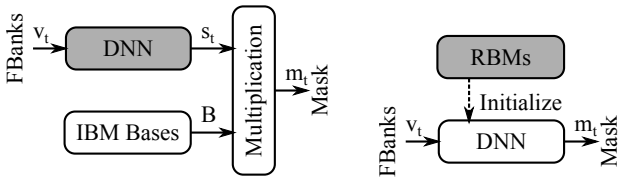
(c)With IBM (d) With ideal state mask

(e) With state dependent mask (f) With DNN based mask

Fig. 2. Spectrograms for the same noisy speech "8055" filtered by different masks. The corresponding clean spectrogram (a) and the original unmasked noisy one (b) are also plotted for comparisons.

Masking noisy signals with IBMs has been shown to substantially improve intelligibility [20] and robustness of ASR systems [21]. They are hence commonly used as the supervision targets for learning mask estimators. Some of the early work uses GMMs [10, 11] or support vector machines (SVMs) [12] to predict IBMs. Recently, DNNs have been adopted for mask predictions in [13]. Totally 26 DNNs are trained on the same feature ensemble to predict the mask values for different frequency channels and a final multi-layer perceptron (MLP) is used to smooth out the DNN predictions. Instead of the discrete IBMs, continuous value based ideal ratio masks are used as targets, which are effectively soft masks. An additional DNN is trained in their system to reconstruct clean features from the masked ones.

In practice, an important consideration is the additional cost brought by the mask estimation, which should not be too taxing on the system. With this in mind, two IBM estimation algorithms that reusing existing models are developed, namely the state dependent (Figure 3(a)) and the DNN based (Figure 3(b)) mask estimations.



(a) State dependent estimator (b) DNN based estimator

Fig. 3. Two proposed mask estimators. The light gray shaded box indicates the model is reused from the original DNN-HMM system.

3. PROPOSED MASK ESTIMATIONS

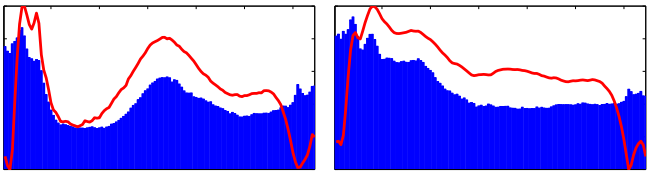
3.1. State Dependent IBM Estimation

Compared to noise, speech is a more structured signal, which is also reflected by the capability of clustering speech features into phonetic clusters for recognition using statistical machine learning methods. Although IBMs depend on both speech and noise due to the local SNR computation, structure information about speech would also be crucial to label out the speech dominate units. For a specific phonetic state cluster, different IBM realizations should share similar structures. Based on this assumption, we propose a state dependent IBM estimation approach. Utilizing the phonetic state clustering of the original recognition system, the IBM vectors for each time frame are grouped according to their corresponding speech feature vectors. There are several advantages of borrowing speech feature based clusters rather than directly clustering IBM vectors. First it saves the clustering cost. Secondly, the feature based clustering would be more robust than the 0-1 based masks and thirdly, it also relieves the clustering process from being constrained by the limited stereo data.

After clustering, a canonical IBM pattern, which will also be referred to as an IBM basis, could then be estimated for each state cluster. In this work, we simply use the mean IBM vector of each state cluster as the basis for that specific phonetic state. It could be interpreted as the expected probability for each T-F unit being marked as speech. With this set of state dependent IBM bases, $\mathbf{B}$, the estimated mask vector for each test feature vector $\mathbf{v}_t$ is computed by $\mathbf{m}_t = \mathbf{B}\mathbf{s}_t$, where $\mathbf{s}_t$ is the basis coefficient vector. We adopt the original DNN posterior probability vectors as $\mathbf{s}_t$. The complete step is also depicted in Figure 3(a). The mask values estimated in this way are in the range of $[0, 1]$ rather than the discrete 0 or 1 as in IBMs. In consideration of the possible errors in estimations for $\mathbf{B}$ and $\mathbf{s}_t$, we take the estimated soft masks directly for spectral masking without binarization. Our approach differs from [22] in the way how these phonetic dependent mask patterns are used. Instead

of using them to refine a current estimation, we directly compose masks from them.

To gain an intuitive understanding on the effectiveness of our simple state dependent mask estimation, two more spectrograms are plotted in Figure 2. For Figure 2(d), we use the ideal posterior vector for $\mathbf{s}_t$ to justify the effectiveness of the IBM bases without worrying about errors in posterior predictions. Due to the use of soft masks, noises could not be completely removed; but compared to Figure 2(b), the speech formant structure becomes much clearer. In practice, we use an existing phonetic DNN to generate $\mathbf{s}_t$ and the spectrogram in Figure 2(e) looks slightly more noisy than Figure 2(d) but still much better than Figure 2(b). Furthermore, in Figure 4 two samples of IBM bases (blue bars) and the corresponding normalized speech spectral envelops are plotted. A strong correlation could be observed, which also validates our early assumption.



(a) The 11th HMM state of "6" (b) The 14th HMM state of "0"

Fig. 4. Comparisons of state dependent bases (blue bars) and speech spectral envelops (red contour) on Aurora2.

3.2. DNN based IBM Estimation

In the state dependent approach, we borrow the posterior information generated from the original DNN. In this section, we revisit the learning procedure for DNNs. A commonly adopted DNN training recipe [23] is to firstly pre-train a stack of restricted Boltzmann machines (RBMs) in an unsupervised way and then discriminatively fine-tune the whole DNN. The learned RBMs are capable of extracting general purpose high-level abstractions that are good representations of the original data and the fine-tuning stage further optimizes them towards a specific task. Hence, if we optimize them for the prediction of IBMs rather than phonetic labels in the fine-tuning stage, the network would then be capable of generating masks for any given inputs (Figure 3(b)).

Comparing to the state dependent approach, instead of reusing the fine-tuned DNN for the basis coefficient computations, we borrow the general purpose RBMs and optimize them for the IBM prediction. The `sigmoid` output units and the mean square error (MSE) criterion are used to replace the `softmax` units and the cross entropy objective. Standard error back-propagation is adopted to optimize the DNN parameters until the reduction of MSE on a validation set is neglectable. With this DNN based IBM estimator, each test utterance is first forwarded to generate the corresponding mask. A new set of FBank features are computed from the masked partial spectrum and forwarded to either the original DNN or the DNN retrained with masked features.

Our DNN based IBM predictor differs from [13] in two major aspects. Firstly, we use one single DNN initialized with existing pre-trained RBMs to directly predict masks from a window of temporal adjacent acoustic frames; while in their approach, a bunch of DNNs are built from beginning to predict the the mask value for each frequency channel and an additional MLP is involved to smooth out the prediction with temporal information captured in the masks. Secondly, the features for mask predictions are different. In our approach, we use the commonly adopted 24D FBank features. However, in [13], an ensemble of different features are concatenated to form the static input coefficients, which include 13D RASTA filtered PLP coefficients, 13D MFCCs, 15D amplitude modulation spectrogram and 6D pitch-based features. Generally speaking, our approach is much easier for an existing DNN-HMM system to incorporate spectral masking.

Similar to previous sections, the mask generated by this DNN based estimator is used to filter the same noisy speech and the resulting spectrogram is illustrated in Figure 2(f). From the visual comparison, this mask generates the most clean-like spectrogram.

4. EXPERIMENTS

In this section, we justify the effectiveness of the proposed two mask estimators in the hybrid DNN-HMM speech recognition systems. Experiments are first carried out on the noisy digit recognition task Aurora2 [17] and then on the medium vocabulary noisy speech recognition task Aurora4 [18].

4.1. Aurora2

The benchmark noisy speech recognition dataset, Aurora2 [17], consists of two sets of training data, one for clean training and the other for multi-style training. Each of them comprises 8,440 utterances and is equally split into 20 subsets. For the multi-style training data, all the utterances in the same subset share the same noise condition and there are totally 4 different noise scenarios (train, babble, car and exhibition hall) at 5 different SNRs (20dB, 15dB, 10dB, 5dB and clean). All the three test sets, A, B and C, are used for evaluation. Set A has the same noises as the multi-style training data and set B has four new noise types, namely restaurant, street, airport and train station. For set C, there are only two noise scenarios (train and street) but with additional channel distortions. For all the three test sets, totally 6 different SNRs are used for evaluation purpose, which have one additional 0dB compared to the training set.

Standard complex back-end GMM-HMM systems are built separately for the clean and multi-style training data using per utterance cepstral mean and variance normalized (CMVN) MFCC features by maximizing the training data likelihood. The 16-state word based HMM and the 5-state silence model are adopted, leading to a total number of 181 HMM states. These GMM-HMM systems are used to generate the per frame DNN training labels. For DNN systems, we use 24D FBank features together with first- and second-order derivatives as inputs. Per utterance CMVN is also adopted for input feature normalization. A consecutive 11 frames of the acoustic features are concatenated as the input to DNNs and we use four 2048D hidden layers. DNNs are trained following [23] with unsupervised RBM pre-training and supervised fine-tuning. No language model is used for this task and an equal probability digit-loop is adopted for decoding only. The open source `Kaldi` toolkit [24] is used for all the experiments.

4.1.1. Spectral Masking

Firstly, experimental results using the clean trained DNN-HMM system are tabulated into the upper half of Table 1. The average 16.1% WER of the baseline is far from humans' expectation. Applying IBMs to spectral features with $LC = 0$, we could have a 3.7% WER, clearly indicating the potential of spectral masking for improving DNNs' noise robustness. With ideal posteriors ("Ideal" in

Table 1. WERs of different masking algorithms on Aurora2.

Style	Masking		WER (%)			
	Train	Test	A	B	C	Avg
clean	-	-	16.9	14.7	17.2	16.1
		IBM	3.6	3.4	4.3	3.7
		Ideal	0.7	0.8	0.8	0.8
		State	15.0	13.6	15.7	14.6
		DNN	5.9	8.4	7.3	7.2
multi	-	-	4.6	**5.3**	**5.1**	**5.0**
		IBM	4.0	4.0	4.5	4.1
		Ideal	0.4	0.5	0.5	0.5
		State	5.1	6.8	6.3	6.0
		DNN	5.3	9.0	6.9	7.1
	IBM	IBM	1.1	1.0	1.2	1.1
	State	State	5.2	7.1	6.5	6.2
	DNN	DNN	**4.1**	6.3	5.4	5.2

Table 1), the state dependent IBM bases could reduce the WER below 1%. Although the ideal posteriors may bring additional information for recognition, the less than 1% WER do imply the potential of this method. However, when the posteriors from the baseline are used for the state dependent mask estimation, we could only obtain a rather small improvement (from 16.1% to 14.6%). The quality of the posteriors is thus crucial to the effectiveness of state dependent masks. Using DNN based estimator, a 7.2% WER could be achieved. The gap between the IBM's performance and our estimators' performances is still quite large. Besides the accuracy of the mask estimators, another probable reason is the mismatch between the clean trained model and the masked features. In our study, the direct masking approach is adopted. The masked features are expected to be different from the clean ones, which could also be observed from the example in Figure 2.

One possible way of solving this problem is the use of multi-style trained model. As with multi-style data, the model has seen much larger variations than the clean data. In "multi" part of Table 1, without any masking, the multi-style trained baseline already has a 5.0% WER. It suggests the importance of data samples from target environments to the good generalization capability of DNNs. Decoding the masked features directly, we could achieve a much better performance than the clean system but worse performance than the "multi" baseline. This suggests that the variations of multi-style data improve DNNs' robustness to masked features but is still not the best fit. Retraining the model with masked multi-style data would be the best solution. From the results in the last three lines of Table 1, the IBM could yield around 1% WER for all the three test sets. The DNN estimator improves from 7.1% to 5.2%; however, it is still a little worse than the baseline's 5.0%. This has to be attributed to the quality of the estimated masks.

Before exploring new mask estimators, if we compare the baseline "multi" DNN and the best one using estimated masks, *i.e.* the one retrained on features filtered by DNN predicted masks, we could find rather different WER breakdowns on each test set. The DNN mask reduces the WER on set A from 4.6% to 4.1%, indicating its effectiveness for known noises. However, for unseen noises in set B, the performances degrades from 5.3% to 6.3%, implying that the DNN mask estimator does not generalize well to unseen noises [25]. This also suggests that using no masks is more preferable than using unreliable ones for the hybrid DNN-HMM systems. Similarly, the performance degrades on set C due to the unseen noises and additional channel distortions. These differences, however, may imply the potential complementariness between these two systems.

4.1.2. Posterior Interpolation

To justify our guess, we simply average the posteriors generated from different systems. The "multi" baseline ("$\mathcal{A}$") is combined with either the system using state dependent mask estimator ("$\mathcal{B}$") or the one with DNN based mask estimator ("$\mathcal{C}$"). Both the previous results for single systems and new ones for two combined systems ("$\mathcal{A}+\mathcal{B}$" and "$\mathcal{A}+\mathcal{C}$") are presented in Table 2. Surprisingly, the simple average between the "multi" baseline and the DNN mask estimator, *i.e.* "$\mathcal{A}+\mathcal{C}$", yields a 4.4% WER, which is much lower than both "$\mathcal{A}$" and "$\mathcal{C}$". This clearly indicates their complementariness to each other.

Table 2. WERs of posterior averaging with the baseline system on Aurora2 for multi-style training.

System	Masking		WER (%)			
	Train	Test	A	B	C	Avg
$\mathcal{A}$	-	-	4.6	5.3	5.1	5.0
$\mathcal{B}$	State	State	5.2	7.1	6.5	6.2
$\mathcal{C}$	DNN	DNN	4.1	6.3	5.4	5.2
$\mathcal{A}+\mathcal{B}$			4.5	5.5	5.2	5.1
$\mathcal{A}+\mathcal{C}$			**3.8**	**5.0**	**4.6**	**4.4**

Additionally, to further understand the performance sensitivity to the interpolation weight, we vary it from 0.0 to 1.0 by 0.1 each time. The final results are illustrated in Figure 5. The performance of the "$\mathcal{A}+\mathcal{C}$" system turns out to be rather robust to the interpolation weight. Any value between 0.1 and 0.9 could yield a WER around 4.5% which is already much better than any of the two models. In other words, adding just 10% of the posteriors from one system to the other is sufficient enough to largely improve both systems' performances. All these clearly indicate the strong complementariness of the two systems. For comparison purpose, we also trained a DNN-HMM system using the combination of the original and masked features and the final average WER is 5.1%, which further confirms that the complementary information comes from the two models rather than the features. Finally, the equal weight interpolation, *i.e.* averaging, yields the best performance of 4.4% WER. To our knowledge, this is among the best performances on the Aurora2 dataset (Table 3).

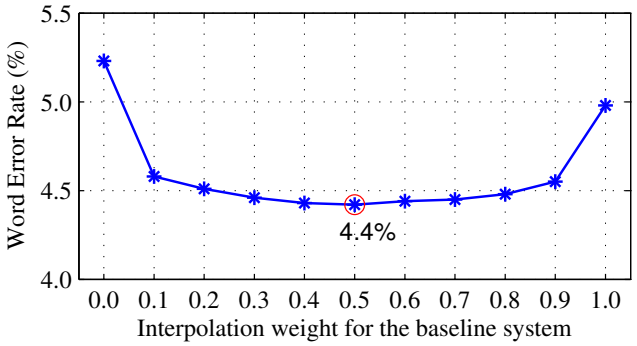


Fig. 5. WER performances of the posterior interpolation between the baselines system and the retrained system using DNN based masks on Aurora2.

Table 3. Reported best WER performances on Aurora2.

System	WER (%)			
	A	B	C	Avg
AFE [26]	6.3	6.7	7.8	6.8
NAT [26]	6.3	6.2	6.1	6.3
DNN+VTSNorm [27]	4.2	5.7	5.3	5.0
ESSEM-MCM [28]	4.4	**4.7**	4.9	4.6
Masking (this work)	**3.8**	5.0	**4.6**	**4.4**

4.2. Aurora4

From the experiments on Aurora2, neither of the two mask estimations perform well on their own. However, the interpolation between the multi-style trained baseline system and the system using masks from the DNN based estimator yields the lowest WER. In this section we thus further investigate the effectiveness of this DNN based mask estimation on a more difficult task - Aurora4 [18].

Aurora4 is a medium vocabulary noisy speech recognition task based on the Wall Street Journal (WSJ0) corpus. Experiments are performed with the 16kHz clean training and multi-style training data respectively. Each of the training set consists of 7138 utterances. For the multi-style training data, one half of the utterances are recorded by the primary Sennheiser microphone and the other half are recorded using one of 18 different secondary microphones. Both halves include a combination of clean speech and speech corrupted by one of six different noises (street traffic, train station, car, babble, restaurant, airport) at 10-20 dB SNR. The evaluation set is derived from WSJ0 5K-word closed vocabulary test set which consists of 330 utterances from 8 speakers. This test set is recorded by the primary microphone and a secondary microphone. These two sets are then each corrupted by the same six noises used in the training set at 5-15 dB SNR, creating a total of 14 test sets. Thus the types of noises are common across training and test sets but the SNRs are not. These 14 test sets can then be grouped into 4 subsets: clean, noisy, clean with channel distortion, noisy with channel distortion, which will be referred to as A, B, C and D, respectively.

Two context-dependent GMM-HMM systems are trained using maximum likelihood estimation on the two training sets and they have 3358 and 3257 senones respectively. The input features are 39D MFCC features including static, first and second order delta features. Per utterance CMVN is performed. These models are used to align the corresponding training data to create senone labels for training the DNN-HMM systems. Decoding is performed with the standard WSJ0 bigram language model.

DNNs are trained using 24D FBank features together with the first and second order derivatives. Utterance level CMVN is adopted. A context window of 11 adjacent frames are used as the DNN inputs and totally 6 hidden layers with 2048 hidden units per layer are trained. The `softmax` output layers have 3358 and 3257 units separately, corresponding to the senones in the GMM-HMM systems.

4.2.1. Spectral Masking

Different masks are first evaluated on the clean trained DNN-HMM system. Unlike Aurora2, the IBM only yields slightly improvement (from 29.2% to 26.2%) and is outperformed by the DNN based mask estimation. One probable explanation is that binary masks may introduce more variations than soft masks used in our DNN based masking. Next the multi-style trained system "multi" is used to evaluate these masks. Comparing the "mutli" baseline to the "clean" one, it performs better under noisy conditions but degrades the performance on clean data, which may due to the high complexity of the Aurora4 task (more than 3,000 senones *v.s.* 181 states on Aurora2 for discrimination). Directly decoding masked features gives rather worse performances, especially for IBM. It could be attributed to the fact that masked features are more similar to clean features rather than noisy ones, which leads to large mismatches between the model and the feature. To address this problem, retraining the DNN with masked features yields improved performances. However, for the DNN based mask estimator, the retrained system has a WER of 13.6% and is still higher than the multi-style baseline's 13.4%, which is similar to what we have observed on Aurora2.

Table 4. WERs of different masking algorithms on Aurora4.

Style	Masking		WER (%)				
	Train	Test	A	B	C	D	Avg
clean	-	-	4.1	22.7	21.7	41.1	29.2
		IBM	4.1	20.0	21.7	36.9	26.2
		DNN	4.1	14.3	21.1	34.8	**22.8**
multi	-	-	5.0	**8.8**	9.0	**20.1**	**13.4**
		IBM	5.0	19.0	9.0	24.1	19.5
		DNN	5.1	12.4	10.4	26.0	17.6
	IBM	IBM	4.9	6.5	8.0	12.2	8.9
	DNN	DNN	**4.7**	9.3	**8.4**	20.3	13.6

4.2.2. Posterior Interpolation

In view of the success in interpolating posteriors of the baseline and the system retrained on masked features, we also investigate this posterior interpolation on Aurora4. Only the interpolation between the multi-style trained DNN and the one retrained on features that are masked with DNN estimated masks is investigated. The interpolation weight for the baseline system also varies from 0.0 to 1.0 and steps by 0.1 each time. The WER results are illustrated in Figure 6. Similarly, a small portion of the posteriors generated by one system would greatly help the other to yield a WER that is much lower than both two, indicating a strong complementariness between these models.

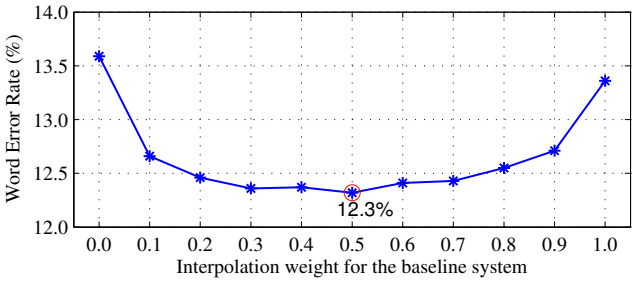


Fig. 6. WER performances of the posterior interpolation between the baselines system and the one retrained on spectrum masked features on Aurora4. The masks are estimated by a DNN.

With an equal weight interpolation, *i.e.* averaging, we could achieve the best WER of 12.3% which is also among the best on Aurora4 (Table 5). Comparing the WER breakdowns among these three systems, our masking approach performs the best on set A and B, which are clean and clean with additive noise. It is exactly the problem that the spectral masking targets to solve - suppressing additive noises. While for data with channel distortions, the masks

Table 5. Reported best WER performances on Aurora4.

System	WER (%)				
	A	B	C	D	Avg
Dropout+NAT [2]	5.4	8.3	**7.6**	18.5	12.4
Masking (this work)	**4.6**	**8.2**	8.4	18.4	12.3
cFDLR [19]	5.1	8.5	8.4	**17.6**	**12.1**

are ill-defined. Moreover, our masking approach operates in spectral feature domain but the "Dropout+NAT" and the "cFDLR" are techniques on models. It is thus possible to combine the best of both worlds, which will be explored in future.

5. CONCLUSIONS

In this paper, we present a low cost approach to integrate the spectral masking technique into the hybrid DNN-HMM system. Two mask estimation methods reusing existing phonetic DNNs are investigated, namely the state dependent and the DNN based mask estimations. Unlike conventional approaches, only FBank features are used in our mask prediction. To further limit the computational costs brought by spectral masking, the data reconstruction process after masking is discarded in the proposed system. Experimental results on Aurora2 and Aurora4 have shown that although the proposed methods fail to beat the multi-style trained DNNs, they do improve the clean trained systems. Most importantly, the average of the posteriors from the baseline and the proposed system using DNN masks could yield WERs of 4.4% on Aurora2 and 12.3% on Aurora4, which are all among the best on these datasets.

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] 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.

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

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

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

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

[7] D.L. Wang, U. Kjems, et al., "Speech intelligibility in background noise with ideal binary time-frequency masking," *The Journal of the Acoustical Society of America*, 2009.

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

[9] B. Raj and R.M. Stern, "Missing-feature approaches in speech recognition," *Signal Processing Magazine*, 2005.

[10] M.L. Seltzer, B. Raj, and R.M. Stern, "A bayesian classifier for spectrographic mask estimation for missing feature speech recognition," *Speech Communication*, 2004.

[11] S. Keronen, H. Kallasjoki, et al., "Mask estimation and imputation methods for missing data speech recognition in a multisource reverberant environment," *Computer Speech & Language*, 2012.

[12] J.F. Gemmeke, Y.J. Wang, et al., "Application of noise robust MDT speech recognition on the SPEECON and speechdat-car databases.," in *Proc. Interspeech*, 2009.

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

[14] S. Keronen, K. Cho, et al., "Gaussian-bernoulli restricted boltzmann machines and automatic feature extraction for noise robust missing data mask estimation," in *Proc. ICASSP*, 2013.

[15] A. Mohamed, G. Hinton, and G. Penn, "Understanding how deep belief networks perform acoustic modelling," in *Proc. ICASSP*, 2012.

[16] N. Jaitly and G. Hinton, "Learning a better representation of speech soundwaves using restricted boltzmann machines," in *Proc. ICASSP*, 2011.

[17] H.G. Hirsch and D. Pearce, "The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *ASR2000-Automatic Speech Recognition: Challenges for the new Millenium ISCA Tutorial and Research Workshop (ITRW)*, 2000.

[18] 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.

[19] 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.

[20] 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.

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

[22] A. Narayanan and D.L. Wang, "Coupling binary masking and robust ASR," in *Proc. ICASSP*, 2013.

[23] G.E. Hinton, "A practical guide to training restricted boltzmann machines," in *Neural Networks: Tricks of the Trade*. 2012.

[24] D. Povey, A. Ghoshal, et al., "The Kaldi speech recognition toolkit," in *Proc. ASRU*, 2011.

[25] 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.

[26] O. Kalinli, M.L. Seltzer, et al., "Noise adaptive training for robust automatic speech recognition," *Audio, Speech, and Language Processing, IEEE Transactions on*, 2010.

[27] 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.

[28] Y. Tsao, J.Y. Li, et al., "Soft margin estimation on improving environment structures for ensemble speaker and speaking environment modeling," in *Proc. IUCS*, 2009.