MULTI-TASK JOINT-LEARNING OF DEEP NEURAL NETWORKS FOR ROBUST SPEECH RECOGNITION

Yanmin Qian[1,2]*, Maofan Yin*[1]*, Yongbin You*[1]*, Kai Yu*[1]

[1] Key Lab. of Shanghai Education Commission for Intelligent Interaction and Cognitive Engineering
Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China
[2] Cambridge University Engineering Department, Cambridge CB2 1PZ, UK
{yanminqian, maofanyin, youyongbin, kai.yu}@sjtu.edu.cn

ABSTRACT

Although deep neural networks (DNNs) have achieved great success in automatic speech recognition (ASR), significant performance degradation still exists in noisy environments. In this paper, a novel *multi-task joint-learning* framework is proposed to address the noise robustness for speech recognition. The architecture integrates two different DNNs, including the regressive denoising DNN and the discriminative recognition DNN, into a complete multi-task structure and all the parameters can be optimized in a real joint-learning mode just from the beginning in model training. In addition, the basic multi-task structure is further explored and reorganized into a more general framework which can get substantial gains. Furthermore, noise adaptive training can also be easily incorporated within this architecture to achieve further performance improvement. Experiments on the Aurora4 task showed that the proposed approach can achieve a WER below 10% without using adaptation or sequence training, a very large and significant (more than 20% relative) improvement over a strong DNN-HMM baseline.

Index Terms— Robust speech recognition, Deep neural network, Feature denoising, Multi-task, Noise aware training

1. INTRODUCTION

Automatic speech recognition (ASR) has come a long way in the last few years, and especially gets great success after the introduction of deep neural network (DNN) based acoustic modelling [1, 2, 3]. Despite the advanced development on ASR systems, noise robustness is still one of the critical issues to make ASR systems widely used in real scenarios. Many technologies [4, 5, 6] have been proposed to handle the problem of mismatch between training and testing, and generally most of these proposed methods can be grouped into two categories: feature denoising and model adaptation.

Feature denoising attempts to remove the corrupting noise from the observations prior to recognition [7]. Model adaptation methods leave the observations unchanged and instead update the model parameters to be more representative of the observed noisy speech [8, 9]. Furthermore, the combination of feature denoising and model adaptation is also performed to get more improved performance. However, most of these algorithms are well designed for the traditional GMM-HMM framework, and there are still no maturely developed approaches to be applied on the DNN-HMM systems.

The recent breakthrough of DNN based acoustic modelling has shown a powerful capacity for ASR [2, 3], and also gotten a promising performance in the noisy scenario [10, 11]. However the noise robustness problem and the mismatch phenomenon still exist in the DNN-HMM systems [12]. Some methods are proposed in the DNN-HMM to improve the noise robustness: In [13], several conventional front-end techniques can still yield gains for DNN-HMM systems for some small tasks, but may lead to a degradation for the large vocabulary tasks [10]. In [14], time-frequency masking and noise adaptive training are used to improve the DNN system in noisy scenarios. More recently, the work in [15, 16, 17, 18] proposed DNN based feature denoising to suppress the noise on the feature level, which is beneficial for later acoustic modelling. Besides, the new DNN structures are also investigated to get better performance [19].

In this paper, inspired by recent work [15, 16, 17] on speech denoising using neural networks, we proposed a novel *multi-task joint-learning* framework to unify feature denoising and acoustic modelling into an integrated model and optimize the parameters in a real joint-learning strategy. Different from the previous work [15, 17] which only used the denoising model as a pre-processor and performed on a feature level independently, the new multi-task architecture combines the regressive denoising DNN and the discriminative recognition DNN into one unified multi-task framework, and all the parameters are optimized considering two different criteria simultaneously from the beginning in training. Compared to the more recent work [18] using a front-end DNN following a back-end DNN, which demands a specific training order on the individual parts, a real joint-learning strategy is implemented on the proposed multi-task model to refine the entire structure simultaneously, and this strategy is relatively more efficient and effective for robust ASR. In addition, the multi-task joint-learning framework is further developed and extended to a more general architecture to get a more improved position.

Moreover subband based noise-aware training is applied to take more advantages of the environmental information to improve the system. The proposed final framework achieves very promising results on the Aurora4 task for all testing cases, and even gets the best published results without any adaptation or sequence training. The

This work was supported by the Program for Professor of Special Appointment (Eastern Scholar) at Shanghai Institutions of Higher Learning, the China NSFC project No. 61222208 and JiangSu NSF project No. 201302060012. Yanmin Qian was partly supported by the EPSRC Programme Grant EP/I031022/1 (Natural Speech Technology). Supporting data for this paper is available at http://www.repository.cam.ac.uk/handle/1810/251280 data repository.

experiments on the unseen noisy scenario also demonstrate the good generalization of the proposed multi-task joint-learning strategy.

The remainder of the paper is organized as follows. In Section 2 a novel multi-task joint-learning framework is proposed for robust speech recognition, including a further developed general architecture and advanced refinement. The experiments and analysis are given in Section 3 and finally draw the conclusions in Section 4.

2. MULTI-TASK JOINT-LEARNING OF DNN

2.1. Multi-task joint-learning for robust speech recognition

Different from the normal DNN using one criterion for optimization, e.g., cross-entropy criterion, the multi-task joint-learning framework usually uses more than one criterion in model training, such as in several previous works: acoustic modelling of triphones and triphonemes [20], multilingual training [21], and phone combining speaker training [22], etc. In this work, we implemented the multi-task joint-learning for the noise robust ASR for the first time. The basic architecture is illustrated as Figure 1, which is also commonly utilized in the other published work [20, 21, 22].

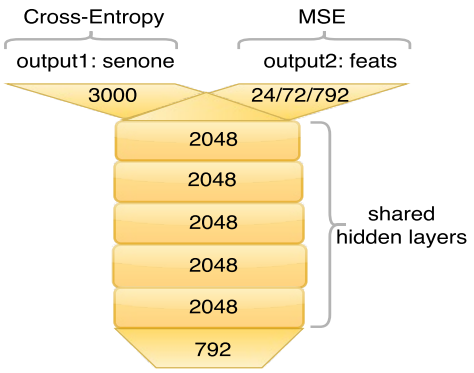

Fig. 1. The traditional multi-task joint-learning framework.

The proposed multi-task joint-learning DNN in Figure 1 has the fully shared hidden layers at the bottom of the model, and two individual task-dependent targets on the outputs: including the regressive denoising DNN and the discriminative recognition DNN with specific criterion for each task. In this multi-task architecture, the denoising task is optimized as a regressive model to predict the clean FBANK features given the input noisy FBANK features[1], and it could be learned using some time-synchronized stereo-data with clean and noisy speech pairs (these pairs can be obtained using synthesizing artificially by corrupting the clean speech utterances with additive noises in various types and SNRs or channel distortions). The predicated target feature could be chosen from three types including the single frame static FBANK (24dim), single frame static plus $\Delta/\Delta\Delta$ FBANK (72dim), or fully context-extended FBANK features (792dim) as the model input. The minimized mean squared error (MMSE) criterion between the DNN outputs and the reference clean features is used for the denoising task. The other recognition task is learned as a discriminative model to predict the state posteriors, same as the normal DNN in ASR, and could use the cross-entropy (CE) criterion. The two objective functions are listed:

$$E_{mse} = \sum_{t=1}^{T} ||\overline{\mathbf{x}}_t - \mathbf{x_t}||_2^2; \; E_{ce} = \sum_{t=1}^{T} \mathbf{D_t} \log(\mathbf{P_t}) \quad (1)$$

where $\overline{\mathbf{x}}_t$ and $\mathbf{x_t}$ are the t^{th} frame vector of estimated and reference clean features respectively in E_{mse}. In E_{ce}, $\mathbf{D_t}$ and $\mathbf{P_t}$ represent the target reference state probabilities and the predicted state posteriors in frame t individually.

In the real optimization, all the parameters of the entire model are jointly learned using both the CE and MMSE criteria from the beginning of training. More specifically, the objective function for the multi-task joint-learning is comprised of two criteria:

$$E(\theta) = E_{ce}(\theta) + \lambda E_{mse}(\theta) \quad (2)$$

where θ represents all the DNN parameters. The $E_{ce}(\theta)$ and $E_{mse}(\theta)$ are the cross-entropy and mean square error function respectively, which are defined in equation (1). λ is a *mixing factor* to balance these two criteria.

Different from the work only using the denoising model as a pre-processor and performing denoising on the feature level independently [15, 17], we could unify feature denoising and acoustic modelling into one integrated multi-task joint-learning framework. In addition, compared to the more recent work using front-end and back-end DNN [18], the proposed new one optimized the parameters in a real joint-learning strategy, and no specific training order is demanded for different model parts. Optimizing the whole parameter set at the same time could take advantages from both two purposes (discrimination & denoising) for all the parameters.

2.2. The more general architecture for multi-task

Normally in the traditional architecture shown as Figure 1, the multi-task joint-learning model will share all the hidden layers for different tasks, and only leave the target outputs individually. It is indeed the most straightforward mode for the multi-task structure, and is relatively easy to design. However, this traditional structure makes all the hidden layers task-independent for all tasks, and no special hidden layers are task-dependent. This may not be very appropriate due to several considerations: (1) one task may be more important than the other, and more task-dependent hidden layers are helpful for that task; (2) some task may be relatively easier to over-training, so fewer hidden layers may be more suitable.

Accordingly a more general architecture is designed for the multi-task joint-learning in Figure 2. All the layers in the new model are divided into three parts, illustrated as the *green, yellow and blue* ones in the figure. The green part represents the shared hidden layers for both tasks, and the other two are the task-dependent layers individually. More specifically, in the work here using the CE and MSE criteria for different tasks, the hidden layer depth for these three parts are indicated as L_{share} (shared hidden layer), L_{ce} (CE-dependent hidden layer) and L_{mse} (MSE-dependent layer) respectively. Comparing Figure 2 with Figure 1, it shows that the traditional framework is actually a special case of the new general one. In the extended general framework, when splitting the different task branches at the last hidden layer and setting both L_{ce} and L_{mse} to 0, it is just the same structure as the traditional one in Figure 1. So the new extended multi-task model is more general, and it could be

[1] $11 * 72 = 792$ dim FBANK features are used in all DNNs in this paper, including the proposed approach and the normal baseline.

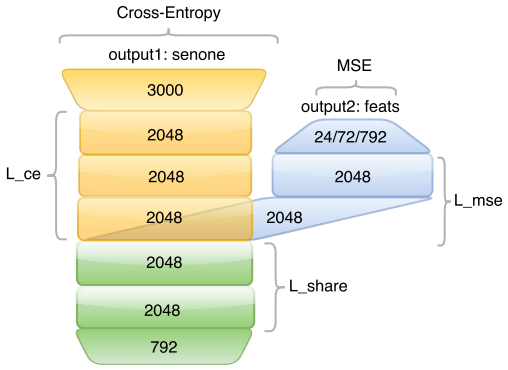

Fig. 2. The extended general multi-task joint-learning.

flexibly adjusted with different configurations for the three parts to get the most optimized performance for robust speech recognition.

No matter which architecture is utilized (the traditional one or the new general one), once finishing the multi-task joint-learning, the denoising-task dependent part is removed, and the remained recognition model could be used as a normal DNN in decoding without any differences from the normal cases. This is also much more flexible and convenient than the preprocessor approaches in [15, 17].

2.3. Subband based Noise-aware Training

The noise information of each utterance is not specifically utilized in the multi-task joint-learning framework described above. To enable this noise awareness, the DNN is fed with the noisy speech features augmented with an estimate of the noise. In this way, the DNN can use additional online noise information to better optimize the model parameters. Also the estimated noise could be regarded as a special noise code for one kind of adaptation. Accordingly in the noise-aware training mode, the input vector of the proposed framework will be changed with the noise estimation appended:

$$\mathbf{V_t} = [\mathbf{Y_{t-\tau}}, ..., \mathbf{Y_{t-1}}, \mathbf{Y_t}, \mathbf{Y_{t+1}}, ..., \mathbf{Y_{t+\tau}}, \mathbf{N_t}] \quad (3)$$

where $\mathbf{Y_t}$ represents the feature vector (FBANK) of the current noisy speech frame t, the context window size is $2 * \tau + 1$, and $\mathbf{N_t}$ is the appended noise code.

Different from the work in [10], we used the subband based noise estimation as the noise code. This noise-estimated vector could encode the noise information on each subband, which makes the noise description much more delicate, and it could retain more useful information for the model training. The subband based noise code for each utterance was computed by averaging the first T frames and fixed for the entire utterance. This is also the first attempt to use the noise-aware training in the multi-task joint-learning DNN. The k subband of noise frame t is:

$$N_t^k = \log\left(\sum_{\mathrm{j=m_k}}^{\mathrm{m_{k+1}-1}} (F_j^2)\right), m_k = \left\lfloor \frac{N_{FFT}}{K} k \right\rfloor \quad (4)$$

where N_{FFT} is the FFT bins number in each frame, F_j is the j^{th} FFT value, and K represents the total subbands number.

3. EXPERIMENTS AND RESULTS

3.1. Experimental setup and baseline systems

To evaluate the proposed multi-task joint-learning framework, a series of experiments were performed on Aurora 4 [23]. Aurora 4 is a medium vocabulary task based on the Wall Street Journal (WSJ0) [24]. It contains 16 kHz speech data in the presence of additive noises and linear convolutional distortions, which were introduced synthetically to clean speech derived from the WSJ0 database. Two training sets were designed for this task: one is a clean-condition training set consisting of 7138 utterances from 83 speakers recorded by the primary Sennheiser microphone, and the other one is the multi-condition training set also comprising 7138 utterances, which is time-synchronized with the clean-condition training set. One half is recorded by the primary Sennheiser microphone and the others are recorded by one of a number of 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 the WSJ0 5K-word closed vocabulary test set which consists of 330 utterances from 8 speakers. This test set was 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 5-15 dB SNR, creating a total of 14 test sets. Notice that the types of noise are common across training and test sets but SNRs of the data 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.

The GMM-HMM system was firstly built to generate the alignments for the DNN-HMM training. This system consisted of context-dependent HMMs with 3K states and 16 Gaussians per state trained using maximum likelihood estimation. The input features were 39-dimensional MFCC features (static plus Δ and $\Delta\Delta$ features) and cepstral mean normalization was performed. After the GMM-HMM building, the forced-alignment was performed to get the senone labels. Decoding was performed with the task-standard WSJ0 bigram language model.

The normal DNN-HMM baseline is constructed for later comparison. It is trained using the 24-dimensional log mel filterbank (FBANK) features with first and second-order derivation, and utterance-level mean normalization was then performed. The input layer was formed from a context window of 11 frames creating an input layer of 792 units for the final DNN. The DNN had 7 hidden layers with 2048 Sigmoid units[2] in each layer and the soft-max output layer had 3K units, corresponding to the senones of the previously GMM-HMM system. The network was initialized using the RBM pre-training [25] and then fine-tuned with a cross-entropy criterion using stochastic gradient descent (SGD) based BP algorithm, with minibatch=128.

For fair comparison, the system in [15] is built, which uses the regressive DNN to denoise the speech, but implemented just as a Pre-processor on the log-power spectra, and then the acoustic model is trained separately. The same configuration is used as in [15], with 257-dim log-power spectra features and 1799-2048-2048-2048-257 architecture (257*7=1799). After the DNN Pre-processor is trained,

[2]The Sigmoid activation function is used in all DNNs in this work

the FBANK features are extracted on the denoised log-power spectra, and normal DNN-HMM system is built as the baseline.

The performances of the baseline and the proposed system in [15] are presented in the top part of Table 1. The DNN-HMM baseline is better than that in [10] due to that the much smaller minibatch 128 is used in our training recipe. And the DNN-PP approach is only slightly better than the baseline, which is consistent (the same 0.2% absolutely better) to the reported results in [15].

Table 1. WER (%) comparisons of the DNN-HMM baseline, the system proposed in [15] using the DNN as a Pre-Processor, and the proposed multi-task joint-learning DNN systems (the regressive denoising DNN models are using different types of target outputs).(**Reg. Out** means the output type of the regressive DNN; **Fbank_Z** denotes the normalized static FBANK feature, **+D_A_Z** denotes the normalized static FBANK feature with Δ and $\Delta\Delta$, and **+11 Frms** denotes the FBANK_D_A_Z feature with 11 context frames extension)

System		A	B	C	D	AVG
DNN-HMM Baseline		4.6	8.2	8.8	18.5	12.4
DNN-PP in [1]		4.1	7.2	7.5	19.4	12.2
MT Joint-Learning		A	B	C	D	AVG
Reg. Out	Fbank_Z	3.9	7.6	7.5	19.1	12.3
	+ D_A_Z	4.0	7.4	6.9	18.4	11.9
	+ 11 Frms	3.9	7.3	6.6	17.3	11.2

3.2. Evaluation of the multi-task joint-learning

3.2.1. Evaluation of the basic multi-task joint-learning

Firstly the multi-task joint-learning is implemented as the traditional structure in Figure 1. The 792-dim FBANK features are utilized as the model input, which is the same as the baseline. There are totally 7 shared hidden layers in the bottom (the same as the baseline) and with the two parallel task outputs on the top. The multi-task DNNs are initialized from the Deep RBMs (Restricted Boltzmann Machines), and then trained as descriptions in Section 2.1. The other training configuration is the same as the baseline: utilizing the same forced-aligned senone labels and finetuning with the SGD (minibatch=128) based BP algorithm.

As stated above, the predicted targets of the regressive DNN could be chosen from several types, such as the static FBANK feature, the static feature with Δ and $\Delta\Delta$, and the feature with long context extension. The results of multi-task joint-learning with different outputs in the denoising DNN are illustrated in the bottom part of Table 1. It shows that the targets of the denoising task are especially important in the proposed architecture. There is nearly no improvement when only using the static feature as the prediction; however the significant WER reduction is obtained when adding the Δ and $\Delta\Delta$ features in the predicted targets. Furthermore it gets another very large improvement when utilizing the long context-extension FBANK feature, i.e., 792 dim. These results demonstrate that making the dynamic features (Δ features as well as the context frames) as the prediction targets is particularly crucial in this multi-task joint-learning framework.

3.2.2. Evaluation of the general multi-task joint-learning

Then we tried to investigate the extended more general multi-task architecture shown as Figure 2. Accordingly various structure configurations are applied to explore the better one for this multi-task training in the robust speech recognition. The setting of L_{share}, L_{ce} and L_{mse} are varied in the experiments, and related results are described in Table 2 and 3.

The denoising task splitting position is firstly investigated. Fixing the total hidden layer number still to 7, the denoising task splitting is pushed to the lower layer. The results are shown in Table 2, and the first line is just the same traditional multi-task structure system of the last line in Table 1. It shows that the system, splitting the denoising task in the lower layer and partly sharing hidden layers ($L_{share} = 3$) in two tasks, is significantly better than the normal fully shared multi-task joint-learning.

Table 2. WER (%) comparisons of the proposed general multi-task joint-learning in different structure configurations, fixing $L_{share} + L_{ce} = 7$, and $L_{mse} = 0$: splitting the denoising task in different positions.

L_{share}	L_{ce}	L_{mse}	A	B	C	D	AVG
7	0	0	3.9	7.3	6.6	17.3	11.2
5	2	0	4.0	7.0	6.9	17.1	11.0
3	4		3.9	7.0	6.6	16.6	10.8
1	6		3.8	7.4	7.2	17.1	11.2

According to the best setting in Table 2 then we fixed the shared hidden layer $L_{share} = 3$ and CE-dependent hidden layer $L_{ce} = 4$, the denoising task dependent hidden layer number is increased from 0 to 3. The results in the top part of Table 3 show that using more non-shared hidden layers (L_{mse}) in the denoising task gets no extra improvement.

Table 3. WER (%) comparisons of the proposed general multi-task joint learning in different structure configurations, fixing $L_{share} = 3$: varying the depth of the task-dependent hidden layer.

L_{share}	L_{ce}	L_{mse}	A	B	C	D	AVG
3	4	0	3.9	7.0	6.6	16.6	10.8
		1	3.8	7.1	6.9	16.6	10.9
		2	4.0	6.9	7.1	16.9	11.0
		3	4.0	7.2	6.6	16.9	11.1
3	7	0	3.9	6.7	6.3	15.4	10.2

Finally, fixing the $L_{share} = 3$ and $L_{mse} = 0$, the depth of the discriminative DNN task (L_{ce}) is increased, and the related results are shown in the last line of Table 3. It shows that more CE-dependent hidden layers are helpful and there is another large improvement when increasing L_{ce} from 4 to 7.

With this structure exploration, it shows that making hidden layers partly shared between tasks and using different task-dependent hidden layers may be helpful in the multi-task joint-learning. The

final new general multi-task architecture obtains more than 2% absolute WER reduction compared to the baseline and also another 1% absolute gain compared to the traditional multi-task structure.

Considering that in the proposed multi-task joint-learning, the model training actually utilized both the multi-condition and clean-condition training set (the clean-condition training set is used as the output targets of the denoising task), and the final best model is relatively larger than the baseline (total 10 hidden layers in the general multi-task DNN ($L_{share} = 3$ & $L_{ce} = 7$ in Table 3) vs. 7 hidden layers in the baseline DNN). For better and fair comparison, several other baselines using more training data and deeper hidden layer are built, and the results are shown in Table 4. The results show that simply increasing the training data and the model depth in the baseline only get a very slight improvement, and this adjustment is not very useful on the normal DNN structure for the noise-robust speech recognition. Compared to the proposed model, the WER decline using the new multi-task joint-learning is significant and very large, which again demonstrates the effectiveness of the proposed new architecture.

Table 4. WER (%) comparisons of the proposed multi-task joint-learning, and the baseline systems using different training data and hidden layer numbers. **M** denotes the multi-condition training set, **C** denotes the clean-condition training set. **Depth** indicates the hidden layer number in the DNN.

Model	Data	Depth	A	B	C	D	AVG
Baseline	M	7L	4.6	8.2	8.8	18.5	12.4
	M+C	7L	4.4	8.1	8.4	18.5	12.3
	M+C	10L	4.3	8.2	7.3	18.3	12.2
Proposed	M+C	10L	3.9	6.7	6.3	15.4	10.2

3.2.3. Evaluation of the enhanced multi-task joint-learning

Finally the subband based noise-aware training, denoted as NAT, is applied on the multi-task joint-learning as described in Section 2.3, and the system comparison is presented in Table 5. It shows that the noise-aware assisted multi-task joint-learning utilizes the information from the environment, which makes the model more robust, and still gets a significant gain in a so strongly improved system.

Table 5. WER (%) comparisons of the proposed multi-task joint-learning with/without NAT.

System	A	B	C	D	AVG
Baseline	4.6	8.2	8.8	18.5	12.4
MT Joint Learning	3.9	6.7	6.3	15.4	10.2
+ NAT	3.8	6.5	6.0	14.5	9.7

To explore the deeper reasons for this refinement, the final average values of the mean squared error (MSE) about the regressive denoising models are plotted for the systems with/without NAT individually[3]. The average MSE values on the Training and CV sets are shown in Figure 3 for the two multi-task systems in Table 5 respectively. It shows that compared to the normal denoising structure, the noise-aware training gets the better average MSE value in the denoising task of the multi-task model, which will finally do the contribution to the other task of discriminative recognition model.

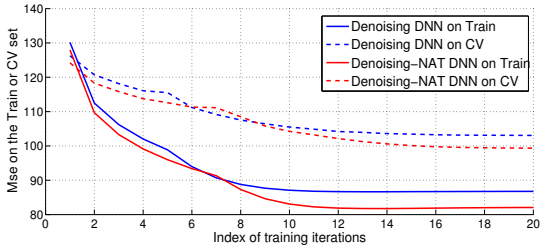

Fig. 3. The average mean squared error (MSE) values of the denoising DNN branch with/without NAT.

Compared to the strong DNN-HMM baseline, the final enhanced multi-task joint-learning DNN framework with NAT gets more than 2.5% absolute (>20% relative) reduction on WER.

3.2.4. Generalization investigation on the proposed framework

To investigate the generalization of the proposed multi-task joint-learning, a more realistic scenario is set up with many unseen noise types other than the noise types in the Aurora 4, which is more similar to the real-world application. The new noisy speech data were synthesized manually by using clean and clean with channel distortion test sets in Aurora 4, which are indicated as SEN and 2ND, and 100 noise types [26] were randomly selected and added to the clean speech at different random SNRs from 5 to 15 dB. Finally 2000 utterances were obtained, with 1000 for each channel. The decoding is performed on this new unseen data using several models, and the results are illustrated in Table 6. It shows that the novel multi-task joint-learning DNNs, trained on specific noise types, are still effective when applied in a totally unseen noisy environment. However as stated in the work in [15], only simply using the regressive DNN independently as the pre-processor leads to the deterioration on system performance when applied in the unseen noise types. The new method still gets about 15% improvement relatively on WER, which demonstrates the good generalization of the proposed multi-task joint-learning architecture for robust ASR.

Table 6. WER (%) comparisons of the baseline and the proposed framework in a more realistic unseen noisy scenario.

System	SEN	2ND	AVG
DNN-HMM Baseline	15.5	24.9	20.2
MT Joint Learning	13.8	21.9	17.8
+ NAT	13.3	20.9	17.1

[3]The average MSE is calculated as E_{mse} in equation (1) and averaged by the total time T, and it is accumulated on all 792 feature dimensions

Table 7. WER (%) comparisons of various systems in the literature to the proposed method on Aurora 4.

System	A	B	C	D	AVG
Best GMM-HMM [9]	5.6	11.0	8.8	17.8	13.4
DNN NAT DP [10]	5.4	8.3	7.6	18.5	12.4
DNN PP [15]	4.5	7.5	7.4	19.3	12.3
Spectral Mask [27]	4.5	7.9	7.5	17.7	11.4
JNAT [14]	4.5	7.4	8.1	16.5	11.1
TVWR Adap [6]	4.4	7.5	7.1	15.6	10.7
Joint FE BE [18]	4.4	6.8	6.4	15.4	10.3
AD OSN LRF [19]	4.0	7.2	6.4	14.5	10.0
MT Joint-learning	3.8	6.5	6.0	14.5	9.7

3.3. The system comparison on Aurora 4

Finally, the results obtained using the proposed approach are compared with several other systems in the literature in Table 7. These systems are representative of the state of the art in acoustic modeling for noise robustness and to the authors' knowledge, are the best published results on Aurora 4. The first system is the best one in the GMM-HMM framework using the VTS and MLLR for environment and speaker adaptations [9], and the systems in the middle block are all the DNN-based systems using various technologies.

As the table shows, the new proposed general multi-task joint-learning DNN with noise-aware training consistently outperforms other works. It achieves a performance below 10% WER without using adaptation or sequence training, and is also much better than the single system in the recent work also using the feature denoising idea but a different framework [18].

4. CONCLUSION AND FUTURE WORK

This paper proposed a novel multi-task joint learning DNN framework for noise robust speech recognition. Different from the previous work of only training the denoising model as a pre-processor or optimizing the parameters independently in a specific order, the new framework unifies the regressive denoising DNN and the discriminative recognition DNN into an integrated and more general multi-task architecture. The new framework can optimize the entire parameter set using two criteria simultaneously from the model training beginning. In addition, the noise-aware training is applied on the subband level and first investigated in the multi-task joint-learning to make use of more environment-related knowledge. With this highly advanced novel structure, the new approach gets more than 2.5% absolute (>20% relative) reduction on WER compared to the strong DNN-HMM baseline on Aurora 4 task. Furthermore, the promising system performance on the unseen noisy scenario also demonstrates the good generalization of the proposed multi-task joint-learning architecture.

In the future, we plan to implement the sequence training [28] on this architecture and also try the other neural function [19] in the multi-task joint-learning model.

5. 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, and Brian 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.

[2] Frank Seide, Gang Li, and Dong Yu, "Conversational speech transcription using context-dependent deep neural networks," in *Proceedings of Interspeech*, 2011, pp. 437–440.

[3] George E Dahl, Dong Yu, Li Deng, and Alex Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 30–42, 2012.

[4] Jinyu Li, Li Deng, Yifan Gong, and Reinhold Haeb-Umbach, "An overview of noise-robust automatic speech recognition," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 22, no. 4, pp. 745–777, 2014.

[5] Yifan Gong, "Speech recognition in noisy environments: A survey," *Speech Communication*, vol. 16, no. 3, pp. 261–291, 1995.

[6] LIU Shilin and Khe Chai Sim, "Joint adaptation and adaptive training of tvwr for robust automatic speech recognition," in *Proceedings of Interspeech*, 2014, pp. 636–640.

[7] Dong Yu, Li Deng, Jasha Droppo, Jian Wu, Yifan Gong, and Alex Acero, "A minimum-mean-square-error noise reduction algorithm on mel-frequency cepstra for robust speech recognition," in *Proceedings of ICASSP*, 2008, pp. 4041–4044.

[8] Michael L Seltzer, Alex Acero, and Kaustubh Kalgaonkar, "Acoustic model adaptation via linear spline interpolation for robust speech recognition," in *Proceedings of ICASSP*, 2010, pp. 4550–4553.

[9] Yongqiang Wang and Mark JF Gales, "Speaker and noise factorization for robust speech recognition," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 20, no. 7, pp. 2149–2158, 2012.

[10] Michael L Seltzer, Dong Yu, and Yongqiang Wang, "An investigation of deep neural networks for noise robust speech recognition," in *Proceedings of ICASSP*, 2013, pp. 7398–7402.

[11] Chao Weng, Dong Yu, Shinji Watanabe, and Biing-Hwang Fred Juang, "Recurrent deep neural networks for robust speech recognition," in *Proceedings of ICASSP*, 2014, pp. 5532–5536.

[12] Yan Huang, Dong Yu, Chaojun Liu, and Yifan Gong, "Multi-accent deep neural network acoustic model with accent-specific top layer using the kld-regularized model adaptation," in *Proceedings of Interspeech*, 2014, pp. 2977–2981.

[13] Bo Li, Yu Tsao, and Khe Chai Sim, "An investigation of spectral restoration algorithms for deep neural networks based noise robust speech recognition," in *Proceedings of Interspeech*, 2013, pp. 3002–3006.

[14] Arun Narayanan and DeLiang Wang, "Joint noise adaptive training for robust automatic speech recognition," in *Proceedings of ICASSP*, 2014, pp. 2504–2508.

[15] Jun Du, Qing Wang, Tian Gao, Yong Xu, Lirong Dai, and Chin-Hui Lee, "Robust speech recognition with speech enhanced deep neural networks," in *Proceedings of Interspeech*, 2014, pp. 616–620.

[16] Andrew L Maas, Quoc V Le, Tyler M O'Neil, Oriol Vinyals, Patrick Nguyen, and Andrew Y Ng, "Recurrent neural networks for noise reduction in robust asr," in *Proceedings of Interspeech*, 2012.

[17] Xue Feng, Yaodong Zhang, and James Glass, "Speech feature denoising and dereverberation via deep autoencoders for noisy reverberant speech recognition," in *Proceedings of ICASSP*, 2014, pp. 1759–1763.

[18] Tian Gao, Jun Du, Lirong Dai, and Chin-Hui Lee, "Joint training of front-end and back-end deep neural networks for robust speech recognition," in *Proceedings of ICASSP*, 2015, pp. 4375–4379.

[19] Steven Rennie, Vaibhava Goel, and Samuel Thomas, "Deep order statistic networks," in *Proceedings of SLT*, 2014, pp. 124–128.

[20] Dongpeng Chen, Brian Mak, Cheung-Chi Leung, and Sunil Sivadas, "Joint acoustic modeling of triphones and trigraphemes by multi-task learning deep neural networks for low-resource speech recognition," in *Proceedings of ICASSP*, 2014, pp. 5592–5596.

[21] Georg Heigold, Vincent Vanhoucke, Alan Senior, Patrick Nguyen, Marc'Aurelio Ranzato, Matthieu Devin, and Jeffrey Dean, "Multilingual acoustic models using distributed deep neural networks," in *Proceedings of ICASSP*, 2013, pp. 8619–8623.

[22] Nanxin Chen, Yanmin Qian, and Kai Yu, "Multi-task learning for text-dependent speaker verification," in *Proceedings of Interspeech*, 2015, pp. 185–189.

[23] 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*, vol. 40, pp. 94, 2002.

[24] Douglas B Paul and Janet M Baker, "The design for the wall street journal-based csr corpus," in *Proceedings of the workshop on Speech and Natural Language*, 1992, pp. 357–362.

[25] Geoffrey Hinton, "A practical guide to training restricted boltzmann machines," *Momentum*, vol. 9, no. 1, pp. 926, 2010.

[26] Guoning Hu, "100 nonspeech environmental sounds," in *http://web.cse.ohio-state.edu/pnl/corpus/HuNonspeech/*, 2004.

[27] Bo Li and Khe Chai Sim, "A spectral masking approach to noise-robust speech recognition using deep neural networks," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 22, no. 8, pp. 1296–1305, 2014.

[28] Karel Veselỳ, Arnab Ghoshal, Lukás Burget, and Daniel Povey, "Sequence-discriminative training of deep neural networks," in *Proceedings of Interspeech*, 2013, pp. 2345–2349.