MULTI-PASS FEATURE ENHANCEMENT BASED ON GENERATIVE-DISCRIMINATIVE HYBRID APPROACH FOR NOISE ROBUST SPEECH RECOGNITION

Masakiyo Fujimoto and Tomohiro Nakatani

NTT Communication Science Laboratories, NTT Corporation, Japan

{fujimoto.masakiyo, nakatani.tomohiro}@lab.ntt.co.jp

ABSTRACT

This paper presents multi-pass feature enhancement technique that consists of three processing passes. In the proposed method, the first pass was described in our previous work, and consists of model-based feature enhancement realized by employing a generative-discriminative hybrid approach with Gaussian mixture models and deep neural networks (DNNs). As an extension of the previous work, the second pass of the proposed method utilizes DNNs retrained with iterative realignment and auxiliary features obtained from intermediate parameters of the first processing pass. In the third pass, we apply unsupervised DNN adaptation and system combination to the results of the second pass. Therefore, the proposed multi-pass technique realizes stepwise improvements in feature enhancement. For CHiME3 task evaluations, the proposed method provided noticeable improvements in noisy speech recognition accuracy compared with results obtained using the previous one-pass feature enhancement technique.

Index Terms— multi-pass feature enhancement, deep neural networks, auxiliary features, unsupervised DNN adaptation

1. INTRODUCTION

The increasing use of mobile devices with speech applications in various environments means that noise robustness has become a crucial problem in relation to automatic speech recognition (ASR). With a view to realizing noise robust ASR, frontend processing including noise suppression and feature enhancement is the simplest way to improve ASR performance in noisy environments. The research and development of frontend processing have been widely pursued using both traditional approaches [1, 2, 3, 4, 5] and recent deep neural network (DNN)-based approaches e.g., a DNN bottleneck feature [6], a denoising autoencoder (DAE) [7, 8, 9], and DNN-based ideal binary masking [10, 11]. Of these various techniques, we have been focusing on feature enhancement, and have already proposed a technique based on a generative-discriminative hybrid approach, which incorporates the benefits of both Gaussian mixture models (GMMs) and DNNs [12].

The aforementioned frontend processing is usually effective for a DNN-hidden Markov model (HMM)-based ASR system trained using clean data; however, it cannot significantly improve to an ASR system with multi-condition training [13]. On the contrary, is sometimes degrades ASR performance with multi-condition training. Thus, this fact suggests that it is difficult to achieve ASR improvement solely with frontend feature enhancement. On the other hand, various training techniques for DNN-based ASR backend have been proposed to overcome this problem including noise adaptive training [14], DNN adaptation [15, 16, 17, 18], and the use of auxiliary features [14, 19, 20, 21]. Among these techniques, DNN training with auxiliary features provides a noticeable improvement in ASR performance. To provide a training scheme with noise awareness, [19] includes the estimated mean vector of the noise in the primary features. This technique is called noise-aware training. [14] investigates the use of various features as alternatives to the noise feature. If we can use multi-channel data, the multi-channel feature will help improve ASR performance [20, 22, 23]. The i-vector input, which is related to speaker awareness, also provides positive results [21, 24, 25].

Inspired by these successful studies of DNN-based ASR backend, in this paper, we explore the effect of incorporating this backend processing in DNN-based frontend processing. Consequently, we propose a multi-pass feature enhancement technique based on the generative-discriminative hybrid approach, which refines the DNNs of the frontend processing at each processing pass. Fig. 1 shows the processing flow of the proposed method. As can be seen in Fig. 1, the proposed method consists of three passes. The first pass is equivalent to our previously proposed generative-discriminative hybrid approach. As an extension of the previous work, the second pass utilizes DNNs retrained with iterative DNN-based realignment and auxiliary features obtained from intermediate parameters of the first pass processing. This processing refines the DNNs for feature enhancement, and provides significant improvements in ASR performance. The thirs pass employs unsupervised DNN adaptation and system combination. The third pass also refines the DNNs by using unsupervised adaptation, and combines the results obtained with the second pass and the unsupervised DNN adaptation of the third pass. In this paper, we employed the simple posterior interpolation-based approach [11] for the system combination. With these processing techniques, we prove that our proposed multi-pass feature enhancement approach outperforms the conventional method and the previously proposed one-pass processing technique.

2. REVIEW OF FEATURE ENHANCEMENT BASED ON GENERATIVE-DISCRIMINATIVE HYBRID APPROACH

This section reviews our previous work, i.e., feature enhancement based on a generative-discriminative hybrid approach with GMMs and DNNs [12]. This part is equivalent to the first pass in Fig. 1.

2.1. GMMs and DNN for feature enhancement

With the proposed method, speech is modeled by using a speaker independent GMM with J Gaussians components in the D-dimensional log mel-filter bank (LMFB) domain. Then, global bias adaptation [26] is applied to the speech GMM, i.e., $\hat{\boldsymbol{\mu}}_{S,j} = \boldsymbol{\mu}_{S,j} + \boldsymbol{b}$, where $\boldsymbol{\mu}_{S,j}$, $\hat{\boldsymbol{\mu}}_{S,j}$ and $\boldsymbol{b}$ denote the mean vector of the speech GMM, the adapted mean vector and the bias vector, respectively. j denotes the Gaussian index of the speech GMM.

The speech DNN with L hidden layers and J output nodes is trained by using the alignment labels of the training data, which give the index sequence of the most likely Gaussian component at the t-th frame [12]. By using these alignment labels, each output node of the

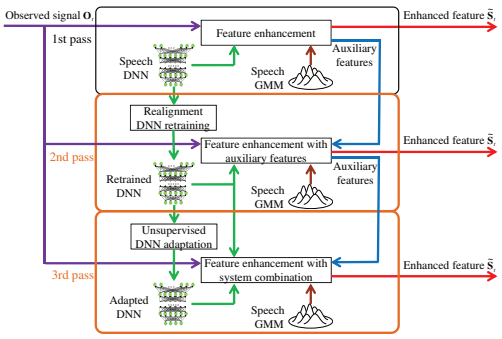

Fig. 1. Processing flow of proposed multi-pass feature enhancement

speech DNN can correspond to each Gaussian component contained in the speech GMM.

With internal processing, the noise is also modeled by using a GMM with K Gaussian components in the LMFB domain.

2.2. Mismatch function and GMM composition

With the LMFB vectors of the speech $\boldsymbol{S}_t$ and the noise $\boldsymbol{N}_t$ at the t-th frame, the LMFB vector of the observed signal $\boldsymbol{O}_t$ is derived by the following mismatch function,

$$\boldsymbol{O}_t = \boldsymbol{S}_t + \mathbf{log}\left(1 + \mathbf{exp}\left(\boldsymbol{N}_t - \boldsymbol{S}_t\right)\right) , \tag{1}$$

where $\mathbf{log}(\cdot)$ and $\mathbf{exp}(\cdot)$ are element-wise operations. Based on this mismatch function, the GMM parameters of the observed signal are obtained by using first-order vector Taylor series (VTS) composition [3] with the GMM parameters of the speech and the noise.

2.3. Computation of discriminative posterior probability

After the GMM composition, the posterior probability w.r.t. the GMM of the observed signal $P_{O,t,j,k}$ is computed. Then, the posterior probability w.r.t. the speech GMM $P_{S,t,j}$ and the posterior probability w.r.t. the noise GMM $P_{N,t,k}$ are given by marginalizing $P_{O,t,j,k}$ as follows:

$$P_{S,t,j} = \sum_k P_{O,t,j,k}, \; P_{N,t,k} = \sum_j P_{O,t,j,k} , \tag{2}$$

where k denotes the Gaussian index of the noise GMM.

With Eq. (2), $P_{O,t,j,k}$ is joint probability of $P_{S,t,j}$ and $P_{N,t,k}$, i.e., $P_{O,t,j,k} = P_{S,t,j} \cdot P_{N,t,k}$. Here, since unsupervised estimation of the noise DNN is currently difficult using only a given observation, we employ the GMM posterior probability $P_{N,t,k}$ for the noise. On the other hand, the posterior probability w.r.t. speech is given by the softmax output of the speech DNN $P^{(DNN)}_{S,t,j}$ instead of the GMM posterior probability $P_{S,t,j}$. Therefore, the discriminative posterior probability $P^{(DNN)}_{O,t,j,k}$ is derived as:

$$P^{(DNN)}_{O,t,j,k} = P^{(DNN)}_{S,t,j} \cdot P_{N,t,k} . \tag{3}$$

2.4. Parameter estimation with MMSE estimates

With the method, the target parameters are bias vector $\boldsymbol{b}$ for speaker adaptation and the parameters of the noise GMM. Each parameter is estimated by using the EM algorithm with the MMSE estimates of the speech $\tilde{\boldsymbol{S}}_t$ and the noise $\tilde{\boldsymbol{N}}_t$ derived as:

$$\tilde{\boldsymbol{S}}_t = \boldsymbol{O}_t + \sum_{j,k} P^{(DNN)}_{O,t,j,k} \cdot \left(\tilde{\boldsymbol{\mu}}_{S,j} - \boldsymbol{\mu}_{O,j,k}\right) \tag{4}$$

$$\tilde{\boldsymbol{N}}_t = \boldsymbol{O}_t + \sum_{j,k} P^{(DNN)}_{O,t,j,k} \cdot \left(\boldsymbol{\mu}_{N,k} - \boldsymbol{\mu}_{O,j,k}\right) , \tag{5}$$

where $\boldsymbol{\mu}_{O,j,k}$ and $\boldsymbol{\mu}_{N,k}$ denote mean vectors of GMMs of the observed signal and the noise, respectively.

With $\tilde{\boldsymbol{S}}_t$ and $\tilde{\boldsymbol{N}}_t$, the target parameters are estimated with maximum likelihood criteria. The accuracies of the MMSE estimates and the target parameters will be mutually improved when the processes are iterated until convergence. The feature enhancement result is obtained as an MMSE estimate $\tilde{\boldsymbol{S}}_t$ at the final iteration. The parameter estimation method is described in further detail in [12, 27].

3. MULTI-PASS FEATURE ENHANCEMENT

This section describes the extensions of the previous work, i.e., the second and third passes indicated in Fig. 1.

3.1. Second pass: DNN retraining and auxiliary features

The second pass consists of DNN retraining with realignment labels and the use of auxiliary features.

3.1.1. DNN-based realignment

Although the speech DNN used in the first pass is trained by using alignment labels, which are obtained with the speech GMM, the use of the DNN will improve the accuracy of the alignment labels thanks to its prominent discriminative ability. Thus, we investigate a way of retraining the speech DNN with the DNN-based realignment labels $Lab^{(DNN)}_t$ given by Eq. (6). Then, further improvement can be expected while iterating DNN-based realignment and retraining.

$$Lab^{(DNN)}_t = \arg\max_j P^{(DNN)}_{S,t,j} \tag{6}$$

3.1.2. Auxiliary features

As mentioned in [19, 21], the use of auxiliary features, e.g., estimated noise features or i-vectors, provides a noticeable improvement in the DNN-based ASR system. Thus, we also investigate a way of introducing auxiliary features into DNN-based feature enhancement. In this paper, the primary features for feature enhancement are the observed signal $\boldsymbol{O}_t$ and their first and second order derivatives. To ensure environmental robustness, we append intermediate feature enhancement parameters, i.e., mean vectors of the speech and the noise, to the primary features as shown in Eqs. (7) and (8). These auxiliary features will offer awareness of both speaker and noise for learning the relationship between the primary feature, the speaker, and the noise.

$$\hat{\boldsymbol{\mu}}_S = \frac{1}{T}\sum_t \tilde{\boldsymbol{S}}_t, \; \hat{\boldsymbol{\mu}}_N = \frac{1}{T}\sum_t \tilde{\boldsymbol{N}}_t \tag{7}$$

$$\boldsymbol{x}_t \triangleq \{\boldsymbol{O}_{t-\tau}, \ldots, \boldsymbol{O}_t, \ldots, \boldsymbol{O}_{t+\tau}, \hat{\boldsymbol{\mu}}_S, \hat{\boldsymbol{\mu}}_N\} , \tag{8}$$

where T, $\boldsymbol{x}_t$, and τ denote the entire number of frames, the extended input vector of DNNs, and the length of the context window, respectively. In Eq. (8), the first and second order derivatives of $\boldsymbol{O}_t$ are omitted for simplification.

3.2. Third pass: DNN adaptation and system combination

After the second pass, the third pass applies unsupervised DNN adaptation and system combination to feature enhancement.

3.2.1. Unsupervised DNN adaptation

Unsupervised DNN adaptation is applied to retrained DNNs of the second pass to mitigate the mismatch and the distortion caused by differences in speakers and noise environments. In this adaptation, we simply retrain the input layer or all the layers of the retrained DNN by using alignment labels obtained from evaluation data. The DNN parameters are updated by iterating a stochastic gradient descent (SGD) with a small learning rate value [15].

3.2.2. System combination with posterior interpolation

In [11], the posterior (softmax output) level combination of two different systems provides positive ASR improvements even with a simple implementation. To obtain further improvement, we also look at a way of combining two feature enhancement passes by interpolating softmax outputs obtained with the retrained DNN (second pass) and the adapted DNN (third pass) as follows:

$$P_{Int,t,j}^{(DNN)} = \alpha \cdot P_{Ret,t,j}^{(DNN)} + (1-\alpha) \cdot P_{Ada,t,j}^{(DNN)}, \qquad (9)$$

where $P_{Int,t,j}^{(DNN)}$, $P_{Ret,t,j}^{(DNN)}$, and $P_{Ada,t,j}^{(DNN)}$ denote the softmax outputs obtained by using interpolation, the retrained DNN, and the adapted DNN, respectively. The interpolation weight α is set at $\alpha = [0, 1]$.

4. EXPERIMENTS

4.1. Experimental setup

ASR evaluations were carried out using the CHiME3 task [28]. The CHiME-3 corpus consists of real six-channel audio data collected in four different environments and additional simulated six-channel data. A tablet device with six microphones was used for audio recording to simulate a situation where a user is talking to the device in daily environments. The considered environments are public transport (BUS), a cafeteria (CAF), a pedestrian area (PED), and a street junction (STR). The corpus includes only read speech, where the sentences to be read were taken from the WSJ0 corpus [29]. The training set comprises 1,600 real and 7,138 simulated utterances, which amount to 18 hours of speech. The development and evaluation sets consist of 3,280 and 2,640 utterances, respectively, each containing real and simulated data at a fifty-fifty split. Both the real and simulated parts were spoken by four different speakers. In this paper, we evaluated several single-channel feature enhancement methods, thus, we used audio data collected by a fifth microphone, which was the closest microphone to the speaker.

The feature parameters for feature enhancement were 24 LMFBs that were extracted by using a Hamming window with a 25 ms frame length and a 10 ms frame shift. The speech GMM was trained by using the clean training data. The GMM had $J = 512$ Gaussian components. The number of Gaussian components of the noise GMM was set at $K = 1, \ldots, 4$. Then, we also trained the speech DNN for the feature enhancement by using the simulated training data. The feature parameters of the DNN were the utterance-wise mean and variance normalized 24 LMFBs and their first and second order derivatives. A context window with $\tau = 5$ was applied to each utterance. We trained five DNNs by changing the number of hidden layers with $L = 1, \ldots, 5$. Each hidden layer had 2,048 nodes and the output layer had $J = 512$ nodes, which correspond to the Gaussian components contained in the speech GMM. The DNNs were initialized by discriminative pre-training with layer-wise back propagation [30]. After the pre-training, the DNN was obtained by fine-tuning with the GMM alignment labels described in Sec. 2.1.

The ASR evaluations were carried out by using a DNN-HMM system. In the training stage, we first built a GMM-HMM system with both the real and simulated training data. The GMM-HMMs were modeled with 3-state tied-mixture triphone HMMs. There were a total of 5,976 HMM states. Each state had 16 Gaussian components. The feature parameters of the GMM-HMMs consisted of 12 PLPs [31], log energy, and their first and second order derivatives. Mean normalization was applied to each utterance. With these GMM-HMMs, we applied HMM state alignment to the training data. After HMM state alignment, we built a DNN-HMM system with discriminative pre-training and fine-tuning. Then, a development set was used for cross validation in fine-tuning. The DNN consisted of five hidden layers. The feature parameters and topology of the hidden layer were the same as those of the DNN for feature enhancement; however, the output layer had 5,976 nodes that corresponded to the HMM states. In all the evaluations reported in this paper, the ASR experiments were performed by using fully composed tri-gram weighted finite state transducers [32] with the DNN-based acoustic model. The evaluation criterion was the word error rate (WER).

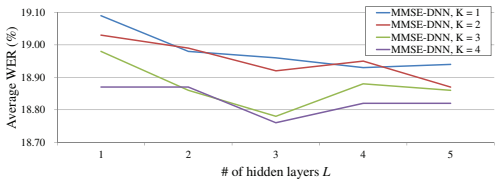

Fig. 2. Average WERs for MMSE-DNN with various model structures

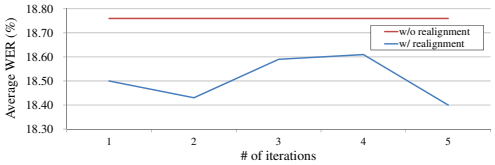

Fig. 3. Average WERs for the retrained DNNs with realignment

4.2. Experimental results of first pass

In this section, we evaluate the first pass in Fig. 1 (MMSE-DNN [12]) by comparison with three conventional methods, i.e., VTS, DAE, and feature enhancement with only the GMM posterior probability (MMSE-GMM [27]).

The adjustable parameters of each method are the numbers of Gaussian components K contained in the noise GMM and hidden layers L of DNN for speech enhancement and DAE. Fig. 2 shows the average WERs for MMSE-DNN with various model structures. As seen in the figure, the WERs tend to improve when both K and L are increased. The parameter values that give the best average WER for each method were $L = 2$ for DAE, $K = 1$ for MMSE-GMM, and $K = 4, L = 3$ for MMSE-DNN, respectively.

Table 1 shows the detailed results of each method and each noise environment. As seen in Table 1, the results obtained with the MMSE-DNN were the best even though other methods deteriorate the WERs from the baseline. As already discussed in [12], these results prove again that discriminative posterior probability must be used for accurate feature enhancement. Thus, the architecture of MMSE-DNN is highly beneficial for improving a DNN-based ASR system with single channel feature enhancement. Hereafter, the MMSE-DNN parameters were fixed at $K = 4$ and $L = 3$.

4.3. Experimental results of second pass

In the second pass, we first evaluated DNN retraining with the DNN realignment labels given by Eq. (6). Fig. 3 shows the average WERs for the retrained DNNs with iterative realignment. In Fig. 3, we can see that the average WER has already been improved at the first realignment. Subsequently, further improvements were obtained while iterating realignment and retraining.

We also evaluated auxiliary features. As mentioned in Sec. 3.1.2, we introduced the mean vectors of the speech $\hat{\boldsymbol{\mu}}_S$ and the noise $\hat{\boldsymbol{\mu}}_N$ to incorporate the awareness of both speaker and noise in feature enhancement. Table 2 shows the detailed results for the second pass. As can be seen, the auxiliary features steadily improved the WERs, in particular, the use of both the speech and noise auxiliary features

Table 1. ASR results for each feature enhancement with WER (%). MMSE-DNN shows results with the first pass of the proposed method.

Feature enhancement	Simulated data					Real data					Total avg.
	Avg.	BUS	CAF	PED	STR	Avg.	BUS	CAF	PED	STR	
Baseline (w/o feature enhancement)	15.04	12.61	17.54	14.70	15.30	23.07	32.49	25.07	18.42	16.29	19.05
VTS	15.59	14.08	18.40	14.33	15.54	23.73	34.17	25.94	18.40	16.40	19.66
DAE ($L = 2$)	15.40	13.65	18.14	14.34	15.45	24.80	36.77	25.25	18.87	18.32	20.10
MMSE-GMM ($K = 1$)	15.16	13.04	17.89	14.36	15.35	23.69	33.65	25.72	18.55	16.83	19.42
MMSE-DNN ($K = 4$, $L = 3$)	**14.65**	**12.42**	**17.37**	**13.80**	**15.02**	**22.87**	**33.03**	**24.65**	**17.81**	**15.99**	**18.76**

Table 2. ASR results for the second pass with WER (%)

Realignment	Auxiliary feature	Simulated data					Real data					Total avg.
		Avg.	BUS	CAF	PED	STR	Avg.	BUS	CAF	PED	STR	
—	—	14.65	12.42	17.37	13.80	15.02	22.87	33.03	24.65	17.81	15.99	18.76
5 iterations	—	14.59	12.66	17.02	13.99	14.68	22.22	32.32	23.76	17.30	15.50	18.40
—	$\hat{\boldsymbol{\mu}}_S$	14.47	12.63	16.90	13.88	**14.48**	22.31	32.68	23.65	17.49	15.41	18.39
—	$\hat{\boldsymbol{\mu}}_N$	14.65	12.85	16.75	14.03	14.96	22.56	32.87	24.39	17.30	15.67	18.60
—	$\hat{\boldsymbol{\mu}}_S, \hat{\boldsymbol{\mu}}_N$	14.47	12.38	16.62	13.86	15.02	22.27	32.49	23.89	17.19	15.52	18.37
5 iterations	$\hat{\boldsymbol{\mu}}_S, \hat{\boldsymbol{\mu}}_N$	**14.23**	**11.99**	**16.60**	**13.58**	14.76	**21.83**	**32.01**	**23.12**	**16.95**	**15.24**	**18.03**

Table 3. ASR results for the third pass with WER (%)

DNN Adaptation	System combination	Simulated data					Real data					Total avg.
		Avg.	BUS	CAF	PED	STR	Avg.	BUS	CAF	PED	STR	
—	—	14.23	11.99	16.60	13.58	14.76	21.83	32.01	**23.12**	**16.95**	15.24	18.03
Input layer, 5 epochs	—	14.05	11.94	16.42	13.49	**14.36**	21.75	**31.56**	23.38	17.02	15.05	17.90
All layers, 5 epochs	—	14.10	12.01	16.27	13.52	14.61	21.77	31.71	23.27	17.06	15.05	17.94
Input layer, 5 epochs	$\alpha = 0.7$	14.07	**11.84**	16.45	**13.43**	14.57	**21.71**	31.69	23.14	17.08	**14.92**	17.89
All layers, 5 epochs	$\alpha = 0.7$	**14.03**	11.86	**16.16**	13.60	14.49	**21.71**	31.65	23.22	17.04	**14.92**	**17.87**

provided noticeable improvements. Moreover, further improvements were obtained by using both DNN retraining and auxiliary features without offsetting each other.

In this paper, to explore the use of the intermediate parameters of an MMSE-DNN, we employed the mean vector of the speech $\hat{\boldsymbol{\mu}}_S$ as the auxiliary feature, which employs speaker awareness. As a representative speaker aware training technique, i-vector is widely used for the speech auxiliary feature, and provides noticeably improved performance [17, 21, 24, 25]. Therefore, we will investigate the use of i-vector together with the auxiliary features used in this paper.

4.4. Experimental results of third pass

Finally, we carried out evaluations with the third pass. The first evaluation was the unsupervised DNN adaptation described in Sec. 3.2.1. DNN adaptation was applied to the input layer or all the layers of the retrained DNN obtained with the second pass, then the total numbers of SGD iterations were set at one, five, and ten epochs. Although only the use of speaker labels is allowed in the evaluation phase of the CHiME3 task, we compared the adaptation scheme by using both speaker and noise labels.

As seen in Figs. 4(a) and 4(b), the results with labels of speaker and/or noise outperform those without any labels. In particular, adaptation with both speaker and noise labels provides noticeable improvements. Therefore, we can confirm that the use of the specific DNNs, which are precisely adapted to individual speaker and noise condition, is crucial factor for DNN-based feature enhancement. However, since CHiME3 task only allows the use of speaker labels, we employed the adapted DNNs with speaker labels, hereafter.

Fig. 5 shows the average WERs for system combination with posterior interpolation. In the figure, the results with $\alpha = 1$ and $\alpha = 0$ are equivalent to the results of the second pass and unsupervised DNN adaptation with five epochs, respectively. As shown in Fig. 5, we can see that system combination provided further improvements if we chose a suitable interpolation weight α.

Table 3 shows the detailed results for the third pass. As seen in Table 3, improvements from the second pass and the third pass were not necessarily significant. However, steady improvements were obtained with unsupervised DNN adaptation and system combination.

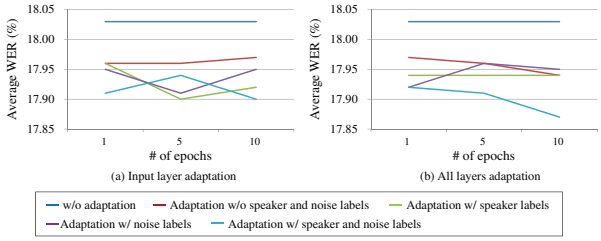

Fig. 4. Average WERs for unsupervised DNN adaptation

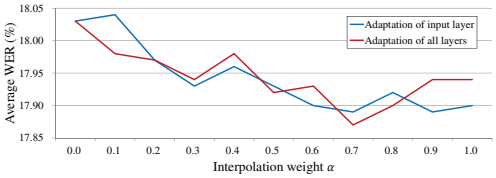

Fig. 5. Average WERs for system combination

5. CONCLUSIONS

This paper described the noticeable improvements realized for DNN-based noise robust ASR by using a multi-pass feature enhancement scheme. The largest impact on the improvement was provided by the second pass, which includes DNN retraining with the realignment labels and auxiliary features obtained by the internal processing of the first pass. However, the steady contributions of the third pass were also indispensable. This work mainly investigated frontend feature enhancement. In future, we plan to introduce various backend processing methods into DNN-based noise robust ASR, and will investigate a way to integrate the processing of the frontend and the backend.

6. REFERENCES

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

[2] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. on ASSP*, vol. 32, pp. 1109–1121, December 1984.

[3] P. J. Moreno, B. Raj, and R. M. Stern, "A vector Taylor series approach for environment-independent speech recognition," in *Proc. of ICASSP '96*, May 1996, vol. II, pp. 733–736.

[4] L. Deng, A. Acero, M. Plumpe, and X. Huang, "Large-vocabulary speech recognition under adverse acoustic environments," in *Proc. of ICSLP '00*, October 2000, pp. 806–809.

[5] ETSI ES 202 050 v.1.1.4, *Speech processing, transmission and quality aspects (STQ), advanced distributed speech recognition; front-end feature extraction algorithm; compression algorithms*, November 2006.

[6] Y. Zhang, E. Chuangsuwanich, and J. Glass, "Extracting deep neural network bottleneck features using low-rank matrix factorization," in *Proc. of ICASSP '14*, May 2014, pp. 185–189.

[7] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder," in *Proc. of Interspeech '13*, August 2013, pp. 436–440.

[8] X. Feng, Y. Zhang, and J. Glass, "Speech feature denoising and dereverberation via deep autoencorders for noisy reverberant speech recognition," in *Proc. of ICASSP '14*, May 2014, pp. 1778–1782.

[9] A. L. Maas, Q. V. Le, T. M. O 'Neil, O. Vinyals, P. Nguyen, and A. Y. Ng, "Recurrent neural networks for noise reduction in robust ASR," in *Proc. of Interspeech '12*, September 2012.

[10] A. Narayanan and D. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," in *Proc. of ICASSP '13*, May 2013, pp. 7092–7096.

[11] B. Li and K. C. Sim, "Improving robustness of deep neural networks via spectral masking for automatic speech recognition," in *Proc. of ASRU '13*, December 2013, pp. 279–284.

[12] M. Fujimoto and T. Nakatani, "Feature enhancement based on generative-discriminative hybrid approach with GMMs and DNNs for noise robust speech recognition," in *Proc. of ICASSP '15*, April 2015, pp. 5019–5023.

[13] M. Seltzer, "Robustness is dead! Long live robustness!," *Keynote of REVERB Workshop*, May 2014, http://reverb2014.dereverberation.com/workshop/slides/mseltzer-reverb2014-keynote-share.pdf.

[14] A. Narayanan and D. Wang, "Joint noise adaptive training for robust automatic speech recognition," in *Proc. of ICASSP'14*, May 2014, pp. 2523–2527.

[15] H. Liao, "Speaker adaptation of context dependent deep neural networks," in *Proc. of ICASSP '13*, May 2013, pp. 7947–7951.

[16] J. Li, J. T. Huang, and Y. Gong, "Factorized adaptation for deep neural network," in *Proc. of ICASSP '14*, May 2014, pp. 5537–5541.

[17] S. Xue, O Abdel-Hamid, H. Jiang, and L Dai, "Direct adaptation of hybrid DNN/HMM model for fast speaker adaptation in LVCSR based on speaker code," in *Proc. of ICASSP '14*, May 2014, pp. 6339–6343.

[18] M. Delcroix, T Yoshioka, A. Ogawa, Y. Kubo, M. Fujimoto, Nobutaka Ito, K. Kinoshita, M. Espi, T. Hori, T. Nakatani, and A. Nakamura, "Linear prediction-based dereverberation with advanced speech enhancement and recognition technologies for the REVERB challenge," in *Proc. of REVERB Workshop*, May 2014.

[19] M. Seltzer, D. Yu, and Y. Wang, "An investigation of deep neural networks for noise robust speech recognition," in *Proc. of ICASSP '13*, May 2013, pp. 7398–7402.

[20] Y. Liu, P Zhang, and T Hain, "Using neural network front-ends on far field multiple microphones based speech recognition," in *Proc. of ICASSP '14*, May 2014, pp. 5542–5546.

[21] G. Saon, H. Soltau, D. Nahamoo, and M Picheny, "Speaker adaptation of neural network acoustic models using i-vectors," in *Proc. of ASRU '2013*, December 2013, pp. 55–59.

[22] A. Schwarz, C. Huemmer, R. Maas, and W. Kellermann, "Spatial diffuseness features for DNN-based speech recognition in noisy and reverberant environments," in *Proc. of ICASSP '15*, April 2015, pp. 4380–4384.

[23] S. Araki, T. Hayashi, M. Delcroix, M. Fujimoto, K Takeda, and T. Nakatani, "Exploring multi-channel features for denoising-autoencoder-based speech enhancement," in *Proc. of ICASSP '15*, April 2015, pp. 116–120.

[24] A. Senior and I. Lopez-Moreno, "Improving DNN speaker independence with i-vector inputs," in *Proc. of ICASSP '14*, May 2014, pp. 225–229.

[25] C. Yu, A. Ogawa, M. Delcroix, T. Yoshioka, T. Nakatani, and J. H. L. Hansen, "Robust i-vector extraction for neural network adaptation in noisy environment," in *Proc. of Interspeech '15*, September 2015, pp. 2854–2857.

[26] M. G. Rahim and B. H. Juang., "Signal bias removal by maximum likelihood estimation for robust telephone speech recognition," *IEEE Trans. on SAP*, vol. 4, no. 1, pp. 19–30, January 1996.

[27] M. Fujimoto and T. Nakatani, "A reliable data selection for model-based noise suppression using unsupervised joint speaker adaptation and noise model estimation," in *Proc. of ICSPCC '12*, August 2012, pp. 4713–4716.

[28] "The 3rd CHiME speech separation and recognition challenge," http://spandh.dcs.shef.ac.uk/chime_challenge/.

[29] J. Garofolo, D. Graff, D. Paul, and D. Pallett, "CSR-I (WSJ0) complete," https://catalog.ldc.upenn.edu/LDC93S6A.

[30] F. Seide, X. Chen, and D. Yu, "Feature engineering in context-dependent deep neural networks for conversational speech transcription," in *Proc. of ASRU '11*, December 2011, pp. 24–29.

[31] H. Hermansky, "Perceptual linear predictive (PLP) analysis for speech," *J. Acoust. Soc. Am.*, vol. 87, no. 4, pp. 1738–1752, April 1990.

[32] T. Hori, C. Hori, Y. Minami, and A. Nakamura, "Efficient WFST-based one-pass decoding with on-the-fly hypothesis rescoring in extremely large vocabulary continuous speech recognition," *IEEE Trans. on ASLP*, vol. 15, no. 4, pp. 1352–1365, May 2007.