Investigation of Monaural Front-End Processing for Robust Speech Recognition Without Retraining or Joint-Training

Zhihao Du* and Xueliang Zhang† and Jiqing Han*
* Harbin Institute of Technology, Harbin, China
E-mail: duzhihao.china@gmail.com, jqhan@hit.edu.cn
† Inner Mongolia University, Hohhot, China
E-mail: cszxl@imu.edu.cn

***Abstract*—There are two effective approaches to improve the performance of an automatic speech recognizer with the front-end processing under noisy condition, one is retraining the acoustic model with the enhanced features, the other is joint-training the acoustic model with the front-end processing model. However, in real life, the automatic speech recognition (ASR) systems are always located in cloud servers but the front-end processing models run locally, which results in the impracticality of the retraining and joint-training strategy for ASR. In this paper, we investigate whether the independent front-end processing can directly improve the performance of a speech recognizer without retraining and joint-training. Three common-used enhancement methods are evaluated in different time-frequency (T-F) domains. Our experiments on CHiME-3 reveal that, with appropriate T-F domains and enhancement methods, the front-end processing can make 35.30% and 11.78% relative word-error-rate (WER) reduction for the Gaussian Mixed Model based (GMM-based) and Deep Neural Network based (DNN-based) recognizer, respectively. For the DNN-based ASR system, we propose using masking-based methods in log-fbank domain to do front-end processing. We find that masking based methods, in general, are better than spectral mapping based methods with respect to WER reduction. In addition, the phases of noisy speech are useless and even harmful to reduce the WER. For generalization capability, the front-end processing can improve the multi-conditional trained ASR system under both matched and unmatched noise condition.**

I. Introduction

Monaural speech enhancement aims at separating speech from the noisy backgrounds by using one microphone. To improve the speech intelligibility and quality, many enhancement methods have been systematically evaluated and successfully utilized. For this purpose, the deep learning based methods significantly improve the enhancement performance [1]. From the perspective of deep learning, training targets, learning machines and input features are three important elements. **1)** Considering the training targets, speech enhancement methods can be divided into three groups, i.e., masking-based, mapping-based and signal approximation based methods. The masking-based methods try to predict a mask computed from premixed noise and clean speech. Wang *et al.* proposed the ideal ratio mask (IRM) in [2], which is frequently used in the supervised speech enhancement. Erdogan *et al.* argued that the phase-sensitive mask (PSM) will lead to higher signal-noise-ratio (SNR) [3]. Williamson *et al.* found that enhancement performance can be further improved by employing the complex ideal ratio mask (cIRM) as the training target [4]. The mapping-based methods try to enhance speech by finding a mapping function between noisy feature and spectrum of the clean speech [5]. The idea of signal approximation (SA) is to train a ratio mask estimator that minimizes the difference between the spectral magnitude of clean speech and that of the estimated speech [6]. **2)** For learning machines, deep neural networks (DNNs) are employed to predict ideal masks [2], [4]. Lu *et al.* used a deep denoising auto-encoder (DDAE) to obtain a clean Mel frequency power spectrogram (fbank) from a noisy one [7]. In [8], [9], convolutional neural networks (CNNs) have been introduced. Besides the feed-forward networks, recurrent networks (RNNs) have also became a popular choice in the speech enhancement community [3]. **3)** As for input features, Wang *et al.* proposed a complementary feature [10] and Chen *et al.* found multi-resolution cochleagram is a better feature in low signal-noise-ratio conditions [11].

Although deep learning based enhancement methods have successfully improved the speech intelligibility and quality, it is not so straight-forward to improve the performance of an ASR system. Compared with human listeners, current ASR systems are more sensitive to the noise interfering and the speech distortion. To improve the robustness of an ASR system, the multi-conditional training strategy is proposed in speech recognition community, which performs acoustic modeling on both clean and noisy utterances. However, This strategy is shown to be effective in matched noise condition but gives an unremarkable performance for the unseen noise [12]. To overcome this issue, there are two strategies introduced to involve the front-end processing in the ASR system. **1)** The first one is using an enhancement model to enhance both training and test sets and retraining the acoustic model with enhanced features [13], [14]. Han *et al.* retrained the acoustic model with features enhanced by a DNN-based spectral feature mapping model. Weninger *et al.* proposed a framework based on Long Short-Term Memory (LSTM) RNNs [15] to enhance the noisy speech in the CHiME-2

and achieved 6.83% improvement by retraining the acoustic model [14]. **2)** The second one is joint-training the front-end enhancement model with the back-end acoustic model [16], [17]. Wang developed a joint training framework and achieve 10.63% relative improvement on the CHiME-2 dataset (task-2), which is the best performance on this dataset [16]. Liu *et al.* proposed an adversarial joint training method and achieve 11.54% relative improvement on the test set of the CHiME-4 challenge [17].

All the above strategies require retraining or joint-training the acoustic model, which are time-consuming and impracticable in real life. Compared with speech enhancement, speech recognition needs handcrafted annotations which make the collection of training data hard and expensive. In real life, the ASR systems are always deployed in cloud servers but the front-end processing models run locally, which results in the impracticality of the retraining and joint-training strategy for ASR. A preferred choice is to train the front-end enhancement model and the back-end acoustic model independently. Therefore, we investigate whether the deep learning based enhancement methods can directly improve the performance of a multi-conditional trained recognizer without retraining or joint-training under the real noisy condition. In this paper, three common-used enhancement methods in different time-frequency (T-F) domains are investigated on the CHiME-3 challenge [18].

The contributions of this paper are as follows:

1) We systematically compare the enhancement methods in different T-F domains, and find that the independent front-end processing can make significant WER reduction for both the GMM-based and DNN-based ASR system. We also find the masking based front-end processing is, in general, more appropriate than the spectral mapping based for an ASR system with respect to WER reduction. And we suggest employing the masking-based method combined with log-fbank domain as the front-end processing for a DNN-based ASR system.
2) We evaluate the effect of noisy phases with respect to WER reduction, and find, with the noisy phases, the front-end processing does not improve the performance of the multi-conditional trained recognizer anymore.
3) We compare the multi-conditional training strategy and the front-end processing under unseen noise condition, and find the front-end processing has stronger generalization capability .

II. Related Work

Without retraining and joint-training, the performance of ASR under the reverberant and simulated noise conditions have been investigated. Wang *et al.* evaluated a masking-based method on the simulated noisy dataset which is derived from Google Voice dataset. For the multi-conditional trained recognizer, Wang's methods gave unremarkable improvement [19]. Li *et al.* proposed an ideal binary mask based enhancement method to improve the noise robustness of a speech recognizer, however, without retraining the acoustic model, Li's method even degraded the WER [20]. Xie *et al.* investigated the effectiveness of the front-end processing under the reverberant condition [21]. There still lacks of a work to systematically examine different enhancement methods and T-F domains for the multi-conditional trained recognizer under real noisy condition.

III. Speech Enhancement Methods

For enhancement methods, ratio masking, direct mapping and signal approximation are three popular choices. All these methods can be performed in different T-F domains, such as power spectrogram, log-fbank domain and etc. In this investigation, we wonder which combination of the enhancement methods and T-F domains is the most appropriate for the multi-conditional trained recognizer. Therefore, we fix our learning machines and focus on the enhancement methods and T-F domains.

A. Enhancement Methods

1) Ratio Masking: The ratio masking-based methods try to learn a mapping function from the noisy features to the T-F masks of the clean speech. The training target of these methods is defined as:

$$\min_{\Phi} \frac{1}{T}\frac{1}{F}\sum_{t=1}^{T}\sum_{f=1}^{F}(RM(t,f) - f_{\Phi}(t,f;\boldsymbol{y}_t))^2 \qquad (1)$$

where T indicates the total frames of the utterance and F represents the number of frequency bins. $\boldsymbol{y}_t$ is the input feature of the enhancement model $f_{\Phi}(\cdot)$, which is extracted from noisy speech at frame t. The enhancement model is parameterized by Φ. $RM(t,f)$ is the desired ratio mask at time t and frequency f which is defined as:

$$RM(t,f) = \frac{S(t,f)}{Y(t,f)} \qquad (2)$$

where $S(t,f)$ and $Y(t,f)$ are the T-F representations of clean and noisy speech, respectively, at each T-F unit. Ratio masks can be defined in different T-F domains, when it comes to short-time Fourier Transform spectral magnitude (FFT domain), such ratio mask is also called *FFT-MASK* in [2] and *SMM* in [1], which is called `fft masking` in this paper. As the ratio masks are not well bounded, we clip them to $[0,1]$ for the training stability as the same manner in [2].

2) Direct Mapping: Mapping-based methods train the enhancement model f_{Φ} to predict the T-F representation of the clean speech from the noisy feature directly. The optimization objective of direct mapping is defined as:

$$\min_{\Phi} \frac{1}{T}\frac{1}{F}\sum_{t=1}^{T}\sum_{f=1}^{F}(S(t,f) - f_{\Phi}(t,f;\boldsymbol{y}_t))^2 \qquad (3)$$

Mapping based methods can be defined in any proper T-F domain, when it is trained to predict the log compressed spectral magnitude, this method is also called FFT-MAG [22] which is called `log-fft mapping` in this paper.

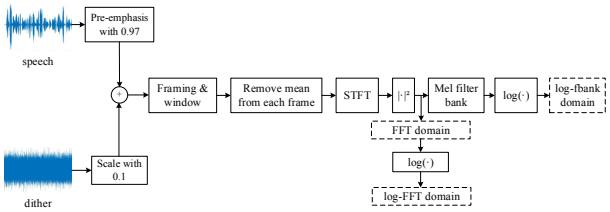

Fig. 1. The extraction pipeline of the investigated domains. The dashed and solid rectangles represent T-F domains and speech processing operations, respectively.

TABLE I
THE INPUT FEATURES, OUTPUT DOMAINS AND ENHANCEMENT METHODS OF THE EVALUATED METHODS.

Evaluated Method	Input Feature	Output Domain	Enhancement Method
log-fbank mapping	log-fbank	log-fbank	mapping
log-fbank SA	log-fbank	log-fbank	SA
log-fbank masking	log-fbank	log-fbank	ratio masking
log-fft mapping [22]	log-FFT	log-FFT	mapping
log-fft SA	log-FFT	log-FFT	SA
log-fft masking	log-FFT	log-FFT	ratio masking
fft mapping	log-FFT	FFT	mapping
fft SA [23]	log-FFT	FFT	SA
fft masking [2]	log-FFT	FFT	ratio masking

3) Signal Approximation: SA-based methods implicitly learn ratio mask from noisy features. Different from the masking-based methods which directly reduce the training loss between the desired mask and the predicted one, SA-based methods reduce the loss between the T-F representations of the target speech and the estimated one. SA-based optimization objective is defined as:

$$\min_{\Phi} \frac{1}{T}\frac{1}{F}\sum_{t=1}^{T}\sum_{f=1}^{F}(S(t,f) - Y(t,f)\odot f_{\Phi}(t,f;\boldsymbol{y}_t))^2 \quad (4)$$

where $\odot$ is the element-wise multiplication. The output of $f_{\Phi}(\boldsymbol{y}_n)$ is restricted to the range $[0,1]$ and bounded as the ratio mask. In [23], the FFT domain is employed to improve Source-to-Distortion-Ratio (SDR), which is called `fft SA` in this paper.

B. T-F Domains

The above enhancement methods can be performed on different T-F domains. In the speech recognition, the log-fbank is found to provide better results compared to mel-frequency cepstral coefficients (MFCC) and log FFT bins [24], so we investigate the enhancement performance on the log-fbank domain. As the log-fbank features can be directly extracted from the FFT domain, we also perform speech enhancement on the FFT domain and its logarithmic counterpart. The extraction pipeline of the investigated domains is demonstrated in Fig.1. The settings (including input features, output domains and enhancement methods) of the evaluated methods are shown in Table I.

IV. EXPERIMENTS

We perform our investigation on the CHiME-3 Challenge [18] which provides multi-channel data for distant-talking automatic speech recognition. There are 7138 clean utterances, 7138 simulated noisy utterances and 1600 real noisy utterances in the training set while there are 410 clean utterance, 1640 simulated noisy utterances and 1640 real noisy utterances in the development set. For the test set, there are 330 clean utterance, 1320 simulated noisy utterances and 1320 real noisy utterances. All the utterances are sampled to 16kHz. We also expand the training set by mixing the clean utterances and the noise records in training set at 0dB, 3dB and 6dB.

A. Speech Recognizer Training

1) Acoustic Model: In the training phase of the acoustic model, we follow the recipe in CHiME-3 challenge to build our baseline. There are two differences between our training and the official training in [18]. First, we train the acoustic model with multi-conditional training strategy (MCT) [12], i.e., we train the GMM-based and DNN-based acoustic model with the clean utterances, the simulated noisy utterances in the fifth channel, the real noisy utterances in the fifth channel and the real close-talking utterances in channel zero while the official training is only performed on simulated and real noisy utterances in fifth channel. The intuition behind this MCT is that the front-end processing tries to reconstruct the clean features, only training the acoustic model with the noisy utterances is obviously unreasonable. Second, we train the acoustic model with log-fbank features instead of MFCC. The log-fbank feature has been widely used in robust speech recognition community [25] and is found to provide better results compared to MFCC and log FFT bins [24]. With the MCT strategy and log-fbank feature, our DNN-based ASR gets lower WER in both development and test dataset than the official GMM-based and DNN-based baseline (seen in table II, III, IV and V). The DNN-based acoustic model is a DNN with 7 hidden layers followed by a sigmoid activation function. After pre-training with the Restricted Boltzmann Machines, the model is fine-tuned by minimizing the cross-entropy loss. We call the official trained acoustic model as `Baseline (official)` and our multi-conditional trained one as `Baseline (MCT)`.

2) Language Model: As the same manners in [18], we employ the WSJ 5k trigram language model and the Kaldi WFST decoder for decoding in all the experiments. Once completing the training, the ASR will be fixed and fed with the enhanced features estimated by the various front-end processing methods.

B. Enhancement Model Training

1) Learning Machine: For the front-end processing, we employ a 4-layer RNN with 512 bidirectional LSTM cells in each layer. In speech enhancement community, RNNs with LSTM cells have been widely employed to leverage the sequential information of speech signals and shown superior performance as compared with DNNs and CNNs [3], [21]. Considering

the value range of the training targets, the last RNN layer is followed by a dense layer with softplus activations for the `fft mapping` methods. And the sigmoid function is employed for the masking-based and the SA-based methods. The linear activation function is used for other enhancement algorithms.

2) Training Target: In the training phase of the front-end models, the input features extracted from the simulated. The real noisy utterances are fed to the models and the corresponding training targets are estimated. All the features are extracted with the window length of 20 ms and the shift length of 10 ms. The hamming window is employed to achieve the periodicity assumption for the Fourier transformation. For the simulated noisy utterances, we can directly extract the clean targets from their corresponding clean counterparts, however, for the real noisy utterances, the training targets are not so straightforward due to the synchronisation issue of the close-talking and the distant-talking microphones. To get the training targets, we align the close-talking and distant-talking utterances by calculating the cross-correlation coefficients. This alignment processing can be performed by solving as following:

$$i^* = \arg\max_i C(X, \overrightarrow{Y}^i) = \arg\max_i \sum_{t=1}^{T} \frac{\sum_{f=1}^{F} X(t,f) \cdot \overrightarrow{Y}^i(t,f)}{\|X(t,\cdot)\|^2 \cdot \|\overrightarrow{Y}^i(t,\cdot)\|^2} \tag{5}$$

where X is the feature of the close-talking utterance, and $\overrightarrow{Y}^i$ represents the i-frame recurrent shifted feature of the distant-talking utterance. We try all valuable i to find the optimal i^* which can maximize $C(X, \overrightarrow{Y}^i)$. For the CHiME-3 challenge, the synchronisation between the close-talking microphone and the distant-talking microphones is only approximate ± 20 ms, so the set of $[-2, -1, 0, 1, 2]$ is enough for i. After the alignment, we use the close-talking utterances and the i^* frame shifted distant-talking utterances to obtain the training targets for the real noisy utterances.

3) Evaluation Method: In evaluating phase, the WER is calculated for the simulated and real noisy utterances in development and test set. Note that the utterances in development set do not appear in the training phase. Different methods are evaluated on the log-fbank domain, log-FFT domain, and FFT domain, respectively. The front-end processing is also performed on the clean and close-talking utterances to find whether it will lead to a degradation on the relatively clean utterances. To evaluate the affect of noisy phases, the ASR is also fed with the synthesized waveforms which are reconstructed from the noisy phases and the estimated spectral magnitudes via the inverse STFT.

V. Results and Discussions

Table II and IV show the WERs of GMM-based and DNN-based ASR on the development set. And the results on the test set are given in Table III and V. The columns with `dt_*` and `et_*` show the results of development and test set, respectively. The WERs of utterances recorded in booth and real noisy environments are given in columns `*_bth` and `*_real`. The columns `*_close` represent the results of close-talking utterances in channel zero and the WERs of simulated noisy speech in fifth channel are shown in columns `*_simu`. The rows marked by `+noisy phases` indicate that we reconstruct waveforms in time domain and extract the ASR features on the waveforms. We investigate this because the ASR systems are always located in cloud servers and need waveforms as input for many real-life scenarios. The average performances of simulated and real noisy utterances are given in the `*_avg` columns.

TABLE II
THE WERS (%) OF GMM-BASED ASR ON DEVELOPMENT SET

Methods	dt_bth	dt_close	dt_simu	dt_real	dt_avg
Baseline (official) [18]	-	-	18.30	18.70	18.50
Baseline (MCT)	5.63	7.52	20.26	21.29	20.78
log-fbank mapping	6.31	7.60	16.87	16.48	16.68
log-fbank SA	5.68	**6.98**	14.99	**15.28**	**15.14**
log-fbank masking	5.74	7.15	15.15	15.54	15.35
log-fft mapping [22]	6.31	8.25	18.99	19.71	19.35
+noisy phases	6.42	8.13	18.00	19.76	18.88
log-fft SA	5.93	7.37	17.40	17.87	17.64
+noisy phases	5.94	7.30	16.56	17.56	17.06
log-fft masking	5.78	7.44	16.66	17.54	17.10
+noisy phases	6.11	7.30	16.27	16.92	16.60
fft mapping	6.03	7.60	17.89	17.06	17.48
+noisy phases	6.36	7.54	17.02	16.94	16.98
fft SA [23]	6.12	7.36	17.12	17.78	17.45
+noisy phases	6.31	7.39	16.85	17.58	17.22
fft masking [2]	**5.56**	7.09	**14.48**	16.19	15.34
+noisy phases	5.99	7.26	14.51	16.39	15.45

TABLE III
THE WERS (%) OF GMM-BASED ASR ON TEST SET

Methods	et_bth	et_close	et_simu	et_real	et_avg
Baseline (official) [18]	-	-	21.50	33.40	27.45
Baseline (MCT)	5.60	14.31	25.00	38.39	31.70
log-fbank mapping	6.39	11.05	18.40	28.56	23.48
log-fbank SA	5.81	**9.99**	16.88	25.87	21.38
log-fbank masking	5.85	10.04	16.98	25.65	21.32
log-fft mapping [22]	6.13	12.14	22.22	30.26	26.24
+noisy phases	6.52	12.20	20.73	30.34	25.54
log-fft SA	6.01	11.07	19.83	28.18	24.01
+noisy phases	5.85	11.04	18.77	27.88	23.33
log-fft masking	5.85	11.56	19.54	27.94	23.74
+noisy phases	5.66	11.45	18.69	27.85	23.27
fft mapping	5.96	11.77	21.21	28.64	24.93
+noisy phases	6.26	11.94	20.45	28.12	24.29
fft SA [23]	5.90	11.58	20.62	29.73	25.18
+noisy phases	5.98	11.69	19.74	29.39	24.57
fft masking [2]	**5.73**	10.22	**16.16**	**24.84**	**20.50**
+noisy phases	5.77	14.38	17.06	27.67	22.37

For the GMM-based ASR (seen in Table II and III), the masking-based method in the FFT domain achieves the best performance, 35.30% relative improvement from 31.70% to 20.50%, on the noisy (including both simulated and real noisy utterances) test set. SA in the log-fbank domain gets the lowest WER on the noisy development set. It seems that the mapping-based method is not a good choice for

TABLE IV
THE WERS (%) OF DNN-BASED ASR ON DEVELOPMENT SET

Methods	dt_bth	dt_close	dt_simu	dt_real	dt_avg
Baseline (official) [18]	-	-	14.30	16.13	15.22
Baseline (MCT)	3.42	4.92	12.68	14.19	13.44
log-fbank mapping	4.04	5.18	13.64	14.40	14.02
log-fbank SA	3.36	4.77	12.30	12.95	12.63
log-fbank masking	**3.36**	**4.73**	**12.08**	**12.70**	**12.39**
log-fft mapping [22]	3.92	5.94	16.57	16.14	16.36
+noisy phases	3.98	5.93	16.49	16.09	16.29
log-fft SA	3.47	5.04	14.84	14.72	14.78
+noisy phases	3.89	5.08	14.66	14.34	14.50
log-fft masking	3.50	5.13	14.59	14.17	14.38
+noisy phases	3.69	5.18	14.49	13.97	14.23
fft mapping	3.82	5.28	15.09	14.60	14.85
+noisy phases	4.13	5.38	14.96	14.16	14.56
fft SA [23]	3.70	5.09	14.51	14.39	14.45
+noisy phases	3.82	5.08	14.50	14.22	14.36
fft masking [2]	3.38	4.93	12.09	13.42	12.76
+noisy phases	3.57	5.02	13.08	14.49	13.79

TABLE V
THE WERS (%) OF DNN-BASED ASR ON TEST SET

Methods	et_bth	et_close	et_simu	et_real	et_avg
Baseline (official) [18]	-	-	21.51	33.43	27.47
Baseline (MCT)	4.03	8.11	15.14	25.44	20.29
log-fbank mapping	4.58	8.36	15.10	26.34	20.72
log-fbank SA	4.09	7.13	13.78	22.73	18.26
log-fbank masking	**3.92**	**7.12**	**13.44**	**22.26**	**17.85**
log-fft mapping [22]	4.52	9.74	19.38	25.87	22.63
+noisy phases	4.76	10.07	19.35	25.90	22.63
log-fft SA	4.11	8.03	17.25	24.78	21.02
+noisy phases	4.20	8.17	16.81	24.20	20.51
log-fft masking	4.09	8.73	17.26	24.89	21.08
+noisy phases	4.30	8.93	17.08	24.70	20.89
fft mapping	4.31	9.23	18.44	25.68	22.06
+noisy phases	4.50	9.47	18.43	25.49	21.96
fft SA [23]	4.15	8.75	17.53	26.59	22.06
+noisy phases	4.48	8.97	17.34	26.09	21.27
fft masking [2]	4.15	7.46	13.65	**22.26**	17.96
+noisy phases	4.09	8.51	15.73	26.32	21.03

the automatic speech recognition purpose. When noisy phase is involved, the masking-based method in the FFT domain degrades significantly on test set. Although the methods in the log-FFT domain are affected slightly by noisy phase, its performances are much worse than the methods in the FFT domain.

For the DNN-based ASR (seen in Table IV and V), the masking-based method in the log-fbank domain is a good choice and achieves 7.78% and 11.78% relative improvement on the noisy development and test set respectively. The masking-based method in the FFT domain gets lower WER than all methods in the log-FFT domain, but it is significantly degraded by noisy phase. The mapping-based front-end processing and the methods in the log-FFT domain do not improve the performance of ASR anymore.

From the results of the GMM-based and DNN-based ASR, we find that the masking-based methods, in general, are more appropriate than the mapping-based and SA-based methods with respect to the WER reduction. The results also reveal that the phases from noisy speech are not helpful to improving the performance of a ASR system.

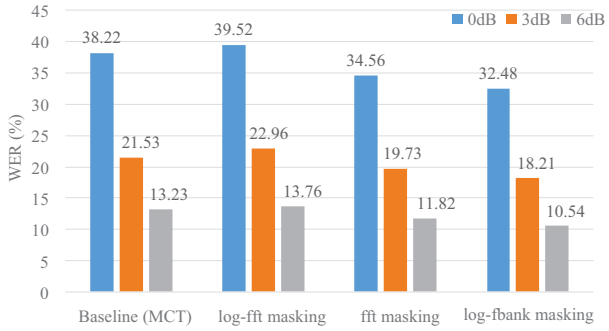

Fig. 2. The WERs (%) of the masking-based methods under the unmatched noise condition with DNN-based ASR.

These front-end processing methods make very little degradation on the relatively clean speech utterances (see the `*_clean` and `*_close` columns). Surprisingly, some methods can even improve the performance of ASR for the close-talking utterances in test set, which is possibly because the close-talking utterances are not very clean but slightly noisy.

To evaluate the generalization capability , we calculate WERs of noisy utterances interfered by babble noise which does not appear in ASR and enhancement model training data. We only evaluate the masking-based methods under the unmatched noise condition as they get better performance under the matched situation. From Table V and Fig.2, we can find the masking-based method in the log-fbank domain achieves the best performance for the unseen babble noise, which also gets the lowest WER under the noise matched condition. We find the ASR with MCT strategy does not generalize well for the unseen noise [12] while the front-end processing efficiently leverages the information of noise and performs better under the unmatched condition.

VI. CONCLUSIONS

In this paper, we investigate the independent front-end processing for robust speech recognition without retraining or joint-training on the CHiME-3 challenge. The masking-based, mapping-based and SA-based methods are evaluated in the log-fbank, log-FFT and FFT domain, respectively. According to this investigation, we suggest employing the masking-based methods as the front-end processing for a multi-conditional trained recognizer rather than the mapping-based methods. For the DNN-based ASR system, we find the ratio masking in the log-fbank domain is a good choice for both matched and unmatched noise conditions. We suggest feeding the recognizer with enhanced features directly and discarding the noisy phases. We find the front-end processing has stronger generalization capability than the multi-conditional training strategy under unseen noise condition.

ACKNOWLEDGMENT

This research was supported by National Science Foundation of China No.61876214, National Key Research and Development Program of China under Grant 2017YFB1002102 and National Natural Science Foundation of China under Grant U1736210.

REFERENCES

[1] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2018.

[2] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 22, no. 12, pp. 1849–1858, 2014.

[3] H. Erdogan, J. R. Hershey, S. Watanabe, and J. Le Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2015, pp. 708–712.

[4] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for monaural speech separation," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 24, no. 3, pp. 483–492, 2016.

[5] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 23, no. 1, pp. 7–19, 2015.

[6] F. Weninger, J. R. Hershey, J. Le Roux, and B. Schuller, "Discriminatively trained recurrent neural networks for single-channel speech separation," in *GlobalSIP, Atlanta, GA, USA*, 2014.

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

[8] L. Hui, M. Cai, C. Guo, L. He, W.-Q. Zhang, and J. Liu, "Convolutional maxout neural networks for speech separation," in *IEEE International Symposium on Signal Processing and Information Technology*. IEEE, 2015, pp. 24–27.

[9] K. Tan, J. Chen, and D. Wang, "Gated residual networks with dilated convolutions," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2018, p. 5.

[10] Y. Wang, K. Han, and D. Wang, "Exploring monaural features for classification-based speech segregation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 2, pp. 270–279, 2013.

[11] J. Chen, Y. Wang, and D. Wang, "A feature study for classification-based speech separation at very low signal-to-noise ratio," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, May 2014, pp. 7039–7043.

[12] F. Li, P. S. Nidadavolu, and H. Hermansky, "A long, deep and wide artificial neural net for robust speech recognition in unknown noise," in *Interspeech*, 2014.

[13] K. Han, Y. He, D. Bagchi, E. Fosler-Lussier, and D. Wang, "Deep neural network based spectral feature mapping for robust speech recognition," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[14] F. Weninger, H. Erdogan, S. Watanabe, E. Vincent, J. Le Roux, J. R. Hershey, and B. Schuller, "Speech enhancement with lstm recurrent neural networks and its application to noise-robust asr," in *International Conference on Latent Variable Analysis and Signal Separation*. Springer, 2015, pp. 91–99.

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

[16] Z.-Q. Wang and D. Wang, "A joint training framework for robust automatic speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 4, pp. 796–806, 2016.

[17] B. Liu, S. Nie, Y. Zhang, D. Ke, S. Liang, and W. Liu, "Boosting noise robustness of acoustic model via deep adversarial training," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

[18] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'chime'speech separation and recognition challenge: Dataset, task and baselines," in *Automatic Speech Recognition and Understanding (ASRU), 2015 IEEE Workshop on*. IEEE, 2015, pp. 504–511.

[19] Y. Wang, A. Misra, and K. K. Chin, "Time-frequency masking for large scale robust speech recognition," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[20] B. Li and K. C. Sim, "Improving robustness of deep neural networks via spectral masking for automatic speech recognition," in *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*. IEEE, 2013, pp. 279–284.

[21] K. Wang, J. Zhang, S. Sun, Y. Wang, F. Xiang, and L. Xie, "Investigating generative adversarial networks based speech dereverberation for robust speech recognition," in *Proc. Interspeech 2018*, 2018, pp. 1581–1585. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2018-1780

[22] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal processing letters*, vol. 21, no. 1, pp. 65–68, 2014.

[23] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Deep learning for monaural speech separation," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 1562–1566.

[24] J. Li, D. Yu, J.-T. Huang, and Y. Gong, "Improving wideband speech recognition using mixed-bandwidth training data in cd-dnn-hmm," in *2012 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2012, pp. 131–136.

[25] J. Li, L. Deng, Y. Gong, and R. Haeb-Umbach, "An overview of noise-robust automatic speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 4, pp. 745–777, 2014.