单声道

Bridging the Gap Between Monaural Speech Enhancement and Recognition with Distortion-Independent Acoustic Modeling

Peidong Wang, *Student Member, IEEE*,
Ke Tan, *Student Member, IEEE*, and DeLiang Wang, *Fellow, IEEE*

***Abstract*—Monaural speech enhancement has made dramatic advances since the introduction of deep learning a few years ago. Although enhanced speech has been demonstrated to have better intelligibility and quality for human listeners, feeding it directly to automatic speech recognition (ASR) systems trained with noisy speech has not produced expected improvements in ASR performance. The lack of an enhancement benefit on recognition, or the gap between monaural speech enhancement and recognition, is often attributed to speech distortions introduced in the enhancement process. In this study, we analyze the distortion problem, compare different acoustic models, and investigate a distortion-independent training scheme for monaural speech recognition. Experimental results suggest that distortion-independent acoustic modeling is able to overcome the distortion problem. Such an acoustic model can also work with speech enhancement models different from the one used during training. Moreover, the models investigated in this paper outperform the previous best system on the CHiME-2 corpus.**

***Index Terms*—speech enhancement, speech recognition, speech distortion, distortion-independent acoustic modeling**

I. Introduction

FORMULATED as a supervised learning problem, speech enhancement has made major progress over the last few years with the use of data driven methods, particularly deep learning. Wang and Wang [25], [26] first introduced deep neural networks (DNNs) to perform time-frequency (T-F) masking for speech enhancement. Lu *et al.* and Xu *et al.* used a deep autoencoder (DAE) or DNN to map from the power spectrum of noisy speech to that of clean speech [11], [30], [31]. Many subsequent studies have been conducted to perform T-F masking or spectral mapping by employing a variety of deep learning models, acoustic features, and training targets [6], [10], [15], [28], [29], [32]. These studies have elevated the performance of speech enhancement by a large margin [21]. DNN-based monaural speech enhancement has improved, for the first time, the intelligibility of noisy speech for human listeners with hearing impairment as well as those with normal hearing [7], [9], [21].

This research was supported in part by two NSF grants (IIS-1409431 and ECCS-1808932) and the Ohio Supercomputer Center.

P. Wang is with the Department of Computer Science and Engineering, The Ohio State University, Columbus, OH 43210 USA (e-mail: wang.7642@osu.edu).

K. Tan is with the Department of Computer Science and Engineering, The Ohio State University, Columbus, OH 43210 USA (e-mail: tan.650@osu.edu).

D. L. Wang is with the Department of Computer Science and Engineering and the Center for Cognitive and Brain Sciences, The Ohio State University, Columbus, OH 43210 USA (e-mail: dwang@cse.ohio-state.edu).

Along with the progress in speech enhancement, researchers have investigated using speech enhancement models as frontends for automatic speech recognition (ASR) systems. Narayanan *et al.* [13], [14] proposed to combine masking-based DNN speech enhancement with speech recognition. With a Gaussian mixture model (GMM) as backend, the enhancement frontend was shown to reduce word error rate (WER) significantly [13]. In a subsequent paper using DNN as backend, the benefit of speech enhancement is mixed, depending on training features [14]. For the acoustic model trained with cepstral features, speech enhancement still helps. With log-Mel features, however, the enhancement frontend causes performance degradation. Du *et al.* [4] applied mapping-based frontends to both GMM and DNN based recognition backends. Their observations are basically in line with those of Narayanan *et al.* The only difference is that their enhancement frontend can yield improvements on clean, noisy, and clean plus channel-mismatched conditions for the DNN acoustic model trained with noisy speech. In the fourth CHiME speech separation and recognition challenge (CHiME-4), Heymann *et al.* [8] noted that the harm of processing artifacts introduced during enhancement may outweigh the benefit brought by noise reduction. Based on these studies as well as our own attempts in applying monaural speech enhancement as a frontend for speech recognition on CHiME-4 corpus, the distortion to speech signals introduced in monaural speech enhancement is a major problem that can render enhancement useless or even harmful for robust ASR.

One way to alleviate the distortion problem is to reduce or eliminate speech distortions in enhancement frontends. Attempts in this direction include a progressive training scheme proposed by Gao *et al.* [5] and a mimic loss proposed by Bagchi *et al.* [1]. Progressive training [5] fine-tunes enhancement models in a multitask manner. Instead of using clean speech as the only target of output layer, they add multiple layers in DNN treating speech with progressively decreased signal-to-noise ratio (SNR) as labels. This way, the enhancement model is trained to reduce noise gradually, as well as the distortion in output layer. The mimic loss based method [1], [16] jointly trains enhancement frontends and recognition backends. It uses senone labels directly as the training target. Experimental results showed that such enhancement frontends can be used with off-the-shelf ASR models in Kaldi [17] on the second CHiME speech separation and recognition corpus (CHiME-2).

In addition to pursuing distortion reduction in speech enhancement models, designing more distortion tolerant acoustic model backends may be another direction. Previous research in speech enhancement field shows that DNNs trained using a variety of noises have the ability to generalize to new noisy conditions [3]. A recent study performed by Narayanan *et al.* [12] investigated the generalization ability of acoustic models trained with various out-of-domain data (noises, bandwidths, codecs, and features). Their observation is that, through large-scale training, such acoustic models perform as well as acoustic models trained with in-domain data.

In this study, we analyze the distortion problem by viewing it as a noise mismatch between training and testing. After comparing five acoustic models, we find that distortion-independent acoustic model can potentially overcome the distortion problem. Experimental results also show that this type of acoustic model can work with speech enhancement frontends different from the one used during training.

The rest of this paper is organized as follows. Section II gives an analysis of the distortion problem, an explanation of distortion-independent acoustic modeling, and a description of utterance-wise recurrent dropout for acoustic model training. Sections III and IV present the experiment setup and results, respectively. We make concluding remarks in Section V.

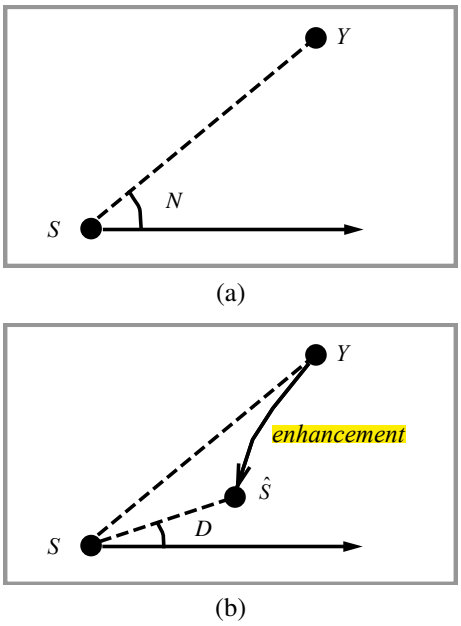

Fig. 1. Illustration of the signal distortion problem. (a) The polar coordinate system. (b) Clean, noisy, and enhanced speech.

II. System Description

A. An Analysis on the Distortion Problem

The distortion in this study refers to the alteration to clean speech signal introduced by speech enhancement that may cause performance degradation in an ASR system. More specifically, this paper tackles with the distortion problem of noise-independent speech enhancement. The input to a speech enhancement system is generated by mixing clean speech with an additive noise, as shown below:

$$y = s + n \tag{1}$$

where y denotes noisy speech, s clean speech, and n an additive noise.

The frequency domain representation of Eq. (1) can be written as (2) below:

$$Y = S + N \tag{2}$$

where Y, S, and N are the spectral representations of noisy speech, clean speech, and additive noise, respectively.

Speech enhancement typically operates on the magnitudes of frequency domain representations. Masking-based models generate a T-F mask, which is then element-wise multiplied with the magnitude of Y,

$$|\hat{S}| = |Y| \otimes M = |S + N| \otimes M \tag{3}$$

where $|\cdot|$ denotes magnitude, $\otimes$ element-wise multiplication, $\hat{S}$ enhanced speech, and M mask.

Depending on the T-F mask definition, M is typically a real-valued matrix with element values ranging from zero to one, e.g. the ideal ratio mask (IRM) [24]. For such masks, (3) can be written as below:

$$|\hat{S}| = |S + N| \otimes M = |S \otimes M + N \otimes M| \tag{4}$$

Thus, we have

$$|\hat{S}| = |S + S \otimes (M - A) + N \otimes M| \tag{5}$$

where A is an all-one matrix.

The distortion for ASR backends can be defined as:

$$D = S \otimes (M - A) + N \otimes M = N \otimes M - S \otimes \overline{M} \tag{6}$$

where $\overline{M}$ denotes the complement of M.

There are two special cases of D. First, if M is an all-one matrix, speech enhancement has no impact on noisy speech. The influence of S on D can also be ignored. Second, let us consider the case when M equals the IRM defined below:

$$IRM = \frac{|S|}{|S| + |N|} \tag{7}$$

In this case, D will be an all-zero matrix, and the distortion problem does not exist.

Other than the two cases above, the influence of S cannot be ignored and the second term in (6) can be viewed as noise residue, which is different from N. Due to this residue, distortion is different from noise N.

Fig. 1 shows the deviation of D from N in an intuitive way. In this figure, spectral representations of different signals are plotted in a polar coordinate system. The center of the coordinate system denotes clean speech S. The distance between clean speech S and noisy speech Y indicates the intensity of noise, and the angle between SY and a predetermined axis indicates noise type N. Fig. 1(a) shows S and its mixture with N, and Fig. 1(b) illustrates the relative positions of Y and enhanced speech $\hat{S}$.

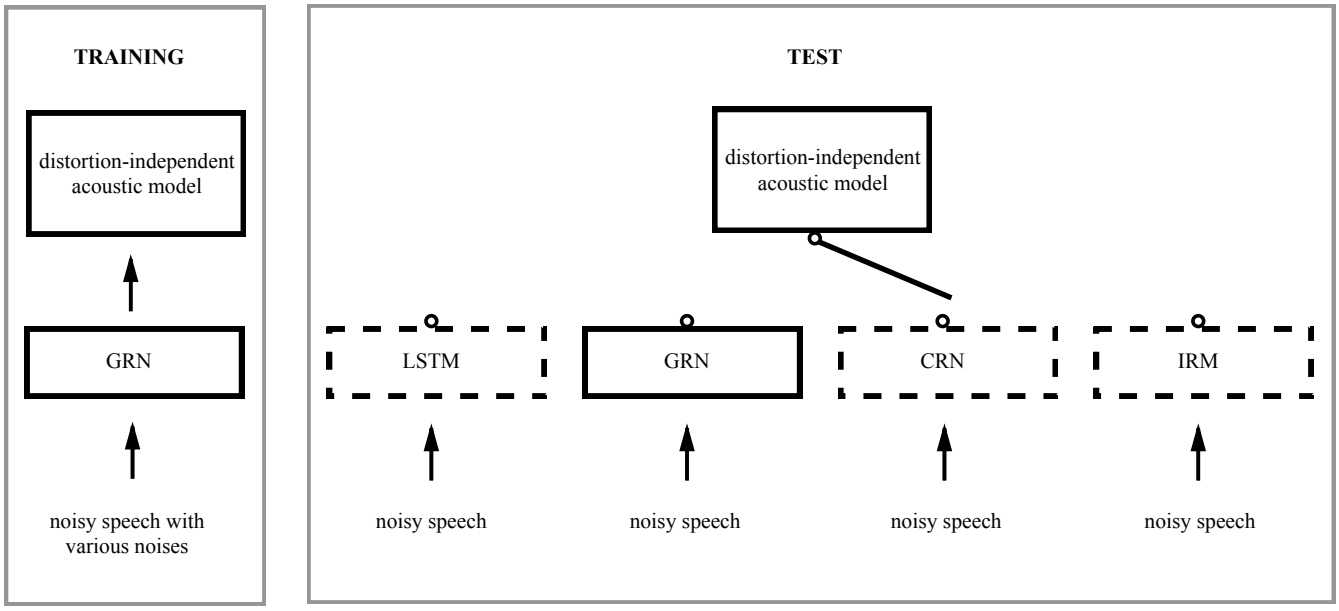

Fig. 2. Illustration of distortion-independent acoustic modeling. See text for the meaning of acronyms.

As is shown in Fig. 1(b), compared with Y, $\hat{S}$ is typically closer to S. This corresponds to the observation that the SNR of enhanced speech is typically higher than that of noisy speech. In fact, many enhancement models are explicitly designed to elevate the SNR.

Along with the shorter distance to S, enhanced speech $\hat{S}$ may deviate from line SY. Such a noise mismatch between Y and $\hat{S}$ may degrade the performance of ASR systems trained only on Y. This may be the main cause of the distortion problem. In fact, for two utterances mixed with the same kind of noise at different SNRs, experimental results suggest that the one with higher SNR typically yield higher recognition performance. Note that, because of the similarity of masking-based and mapping-based speech enhancement in terms of distortion, the analysis above is expected to be valid for mapping-based systems as well.

B. Distortion-Independent Acoustic Modeling

For ASR backends trained on noisy speech and evaluated on enhanced speech, the input data for training and evaluation can be expressed as (8) and (9), respectively,

$$|Y_{tr}| = |S_{tr} + N_{tr}| \tag{8}$$

$$|\hat{S}_{eval}| = |S_{eval} + D_{eval}| \tag{9}$$

where $D_{eval} = N_{eval} \otimes M_{eval} - S_{eval} \otimes \overline{M_{eval}}$. Subscripts tr and $eval$ denote training and evaluation, respectively. Y_{tr}, S_{tr}, and N_{tr} are the spectral representations of noisy speech, clean speech, and additive noise in training, respectively. $\hat{S}_{eval}$ is the enhanced speech in evaluation. D_{eval} denotes the distortion in enhanced speech and M_{eval} the T-F mask in evaluation.

Based on our analysis in the previous subsection, the mismatch between N_{tr} and D_{eval} is the cause of the distortion problem. In speech recognition corpora such as Aurora and CHiME series, only a limited number of noises are provided for training. In addition, noise types are shared between training and evaluation on these corpora. ASR systems trained with such noisy speech may not perform well on enhanced speech, which contains mismatched interference D_{eval}. Moreover, same noise types between training and evaluation give an advantage to unenhanced evaluation data. This is likely a main reason why speech enhancement does not improve recognition performance on these tasks.

To alleviate the distortion problem, N_{tr} can be modified in two ways. If we view D_{eval} as a special type of additive noise, a straightforward way is to increase the scope of N_{tr}. Since this strategy typically uses a large variety of additive noises to train acoustic models, we denote it noise-independent acoustic modeling. An advantage of noise-independent training is that its efficacy is not influenced by speech enhancement frontends. This acoustic modeling strategy, however, does not account for the fact that additive noises may differ significantly from distortions. Another strategy to alleviate the distortion problem is to train the acoustic model directly with enhanced speech, i.e.

$$|\hat{S}_{tr}| = |S_{tr} + D_{tr}| \tag{10}$$

where $\hat{S}_{tr}$ denotes enhanced training speech and D_{tr} refers to the distortion in it.

We investigate a distortion-independent acoustic modeling method based on (10). The training set consists of a large variety of enhanced speech generated by a single well-trained speech enhancement frontend. The input to the speech enhancement model is noisy speech with various types of additive noise. An advantage of distortion-independent acoustic modeling is that D_{tr} in enhanced training speech is similar to D_{eval} during evaluation. The main concern is its generalization

ability to other speech enhancement frontends. Since most supervised speech enhancement models can be viewed as nonlinear mapping from noisy speech to clean speech, distortion-independent acoustic model may be able to work with speech enhancement frontends different from the frontend used for training.

Fig. 2 illustrates distortion-independent acoustic modeling. The left diagram depicts the training stage and the right one testing. In the right diagram, speech enhancement blocks with dashed lines denote those not used during training. In this study, we evaluate three existing speech enhancement models: gated residual network (GRN) [18], LSTM [2], and convolutional recurrent network (CRN) [19]. We also add the IRM as another enhancement frontend. The switch in the right diagram denotes the coupling between a distortion-independent acoustic model and various enhancement frontends.

C. Types of Acoustic Models

In addition to noise-independent and distortion-independent acoustic models, we investigate three other types of acoustic models: clean, noise-dependent, and noise-mismatched. The clean acoustic model is trained using clean speech. In corpora containing both additive noise and reverberation, clean refers to reverberant speech without noise. The noise-dependent acoustic model is trained using only one type of noise and is tested on the same type of noise. This experimental setup represents typical robust speech recognition evaluations. The noise-mismatched acoustic model also uses a single type of noise during training, but it is tested on noises different from those for training.

D. Utterance-Wise Recurrent Dropout for Acoustic Model Training

Utterance-wise recurrent dropout has been shown to be effective for acoustic model training on the CHiME-4 corpus [23]. A typical LSTM layer is described in Eqs. (11), (12), and (13) below:

$$\begin{pmatrix} \mathbf{i}_t \\ \mathbf{f}_t \\ \mathbf{o}_t \\ \mathbf{g}_t \end{pmatrix} = \begin{pmatrix} \sigma(\mathbf{W}_i \mathbf{x}_t + \mathbf{U}_i \mathbf{h}_{t-1} + \mathbf{b}_i) \\ \sigma(\mathbf{W}_f \mathbf{x}_t + \mathbf{U}_f \mathbf{h}_{t-1} + \mathbf{b}_f) \\ \sigma(\mathbf{W}_o \mathbf{x}_t + \mathbf{U}_o \mathbf{h}_{t-1} + \mathbf{b}_o) \\ f(\mathbf{W}_g \mathbf{x}_t + \mathbf{U}_g \mathbf{h}_{t-1} + \mathbf{b}_g) \end{pmatrix} \tag{11}$$

$$\mathbf{c}_t = \mathbf{f}_t \otimes \mathbf{c}_{t-1} + \mathbf{i}_t \otimes \mathbf{g}_t \tag{12}$$

$$\mathbf{h}_t = \mathbf{o}_t \otimes f(\mathbf{c}_t) \tag{13}$$

The dropout method can be expressed as follows:

$$\begin{pmatrix} \mathbf{i}_t \\ \mathbf{f}_t \\ \mathbf{o}_t \\ \mathbf{g}_t \end{pmatrix} = \begin{pmatrix} \sigma(\mathbf{W}_i d_{xit}(\mathbf{x}_t) + \mathbf{U}_i d_{hi}(\mathbf{h}_{t-1}) + \mathbf{b}_i) \\ \sigma(\mathbf{W}_f d_{xft}(\mathbf{x}_t) + \mathbf{U}_f d_{hf}(\mathbf{h}_{t-1}) + \mathbf{b}_f) \\ \sigma(\mathbf{W}_o d_{xot}(\mathbf{x}_t) + \mathbf{U}_o d_{ho}(\mathbf{h}_{t-1}) + \mathbf{b}_o) \\ f(\mathbf{W}_g d_{xgt}(\mathbf{x}_t) + \mathbf{U}_g d_{hg}(\mathbf{h}_{t-1}) + \mathbf{b}_g) \end{pmatrix} \tag{14}$$

where $\mathbf{i}_t$, $\mathbf{f}_t$, and $\mathbf{o}_t$ are the input, forget, and output gates at step t; $\mathbf{g}_t$ is the vector of cell updates and $\mathbf{c}_t$ denotes updated cell vector; $\mathbf{c}_t$ is used to update hidden state $\mathbf{h}_t$; σ is a sigmoid function and f is typically chosen to be $tanh$. $\mathbf{W}$ and $\mathbf{U}$ are the weight matrices for the input vector $\mathbf{x}_t$ and hidden vector $\mathbf{h}_{t-1}$, respectively. $\mathbf{b}$ denotes the bias term. The dropout function is denoted as $d()$. Subscripts x and h refer to the two corresponding feature vectors and i, f, o, g correspond to the four LSTM components. Dropout functions with subscript t are conventional frame-wise dropout, and those without t are recurrent, i.e. they use the same dropout mask at different time steps.

Utterance-wise recurrent dropout is designed to be both recurrent and have little temporal information loss. Four independently sampled utterance-wise masks are applied to $\mathbf{h}_{t-1}$. For the dropout on $\mathbf{x}_t$, we opt for a conventional frame-wise method since utterance-wise dropout may completely lose the information in some feature dimensions.

III. Experimental Setup

A. Datasets

We use two corpora in our experiments. One of them is designed specially for this study and the other one follows the official CHiME-2 recipe.

1) WSJ: We compose a corpus by mixing clean speech in WSJ with additive noise. Although such simulated corpora are not commonly used in speech recognition, they are common in speech enhancement [3], [18].

Training sets for the five acoustic models are designed in the following way. For the clean acoustic model, the clean utterances in the original WSJ corpus are used directly. The noise-dependent acoustic model has two instances, each corresponding to a different noise. The noise-mismatched acoustic model also has two instances, but it differs from the noise-dependent acoustic model in that its training and testing noises are mismatched. The training sets for the noise-dependent and noise-mismatched acoustic models are the same. It contains 7138 utterances generated by mixing clean utterances with a training noise (ADTbabble or ADTcafeteria1) at SNRs randomly chosen from {9dB, 6dB, 3dB, 0dB, -3dB, -6dB}. ADTbabble and ADTcafeteria1 (available at http://www.auditec.com) are commonly used in speech enhancement tasks [3], [18]. For the noise-independent acoustic model, the training set is generated by adding noise segments from a 10000 noise database (available at https://www.soundideas.com) to clean utterances at SNRs randomly chosen from the above six levels. The size of the noise-independent training set is 157036, 22 times that of the clean training set. The distortion-independent acoustic model is trained using GRN enhanced speech. GRN takes as input the noisy speech used for noise-independent acoustic model training. The distortion-independent training set thus also contains 157036 utterances.

A validation set is shared among the five acoustic models. It contains 1206 clean utterances from 10 speakers different from those used in training sets. Note that clean utterances are used directly in the validation set, avoiding biases to any specific noise.

The five acoustic models also share the same test set. It consists of 330 noisy utterances for each of the two test noises

(ADTbabble and ADTcafeteria1) and at each of the six SNRs (i.e. {9dB, 6dB, 3dB, 0dB, -3dB, -6dB}). The total number of utterances is 3960. These utterances are from 12 speakers different from those in the training and validation sets.

Note that although ASR backends and enhancement frontends both use the 10k noise database, their actual noise segments are different. First, ASR backends only use the first halves of noises, and enhancement frontends the second halves. Second, noise segments are randomly selected for recognition and enhancement.

2) CHiME-2: CHiME-2 is a commonly used corpus for robust speech recognition. Different from WSJ, utterances in CHiME-2 contain room reverberation. We treat reverberant speech in CHiME-2 as clean speech.

Training sets for the five acoustic models are designed based on the official recipe of the CHiME-2 challenge. The reverberant acoustic model is trained using reverberant utterances. Since each recording in CHiME-2 has two channels, we apply an average operation to get the corresponding monaural utterance. The noise-dependent acoustic model exactly follows the CHiME-2 recipe. The noise-mismatched acoustic model tests the noise-dependent acoustic model on ADT noises (ADTbabble and ADTcafeteria1) rather than the CHiME-2 noises. Due to the limited number of noises provided in the CHiME-2 corpus, the noise-independent acoustic model is trained with additional noises from the 10k noise database. We mix reverberant utterances with noise segments. The SNR levels are the same as those for WSJ. The noise-independent training set contains 157036 utterances in total. For the distortion-independent acoustic model, the training set consists of 157036 utterances enhanced by GRN.

For the noise-dependent acoustic model, we apply a validation set consisting of noisy utterances. For the other four acoustic models, we use reverberant utterances.

In addition to the official CHiME-2 test set, we generate two other test sets containing ADT noises. The average results on ADT noises are reported in this paper.

Due to reverberation, speech enhancement models for CHiME-2 are trained to map from reverberant-noisy speech to reverberant speech. The training data for speech enhancement models are generated similarly to those for the noise-independent acoustic model.

B. Implementation Details

We use a wide residual bidirectional LSTM network (WRBN) as the DNN architecture of acoustic models [8], [22], [23]. For speech enhancement frontends, we adopt three models as illustrated in Fig. 2. GRN is the main frontend in our experiments and is used to generates both training data and test data. Two other speech enhancement models, LSTM and CRN, generate additional test data for distortion-independent acoustic modeling. These three frontends use different training targets. GRN applies the phase sensitive mask (PSM), LSTM uses the IRM, and CRN is mapping based.

We couple enhancement frontends and ASR backends with enhanced waveforms. The preprocessing steps for the enhancement frontends include windowing and Fourier transform. We apply the Hamming window with window width 20ms and shift 10ms. The windowed waveform signals are then converted to 320-dimensional short-time Fourier transform (STFT) features. Speech enhancement models take as input the STFT magnitudes and generate masks or enhanced magnitudes. We combine enhanced magnitudes with the phase of noisy speech to resynthesize enhanced waveform signals. As for the feature preprocessing for ASR backends, we make modifications to the recipe in Kaldi and our previous experiments [22], [23]. In order to avoid manually added interferences to enhanced speech, we skip most of preprocessing steps, including pre-emphasizing, dithering, and direct currency offset removal. Similar to speech enhancement frontends, we extract spectral features from enhanced waveform signals by applying the Hamming window and performing STFT. One difference is that STFT features for speech recognition have 512 dimensions. We then apply Mel filters to STFT magnitudes to generate Mel frequency features. The dimension of Mel features is 80. In order to avoid underflow, we add a small value e^{-40} to Mel features and apply logarithm to the summation. The delta and delta-delta of log-Mel features are then generated, tripling the size of feature dimensionality. We calculate the mean value along time for each utterance and subtract it from the features. ASR backends take as input the normalized features and generate log posterior probabilities for senones. There are 1965 senones in our experiments. Subtracting log priors from log posteriors, we feed log likelihoods to the decoder in CHiME-2 to generate transcriptions.

In training the noise-independent and distortion-independent acoustic models, we monitor validation results after every 7138 utterances. This technique is commonly used in speech and language processing experiments.

1) WSJ: During training, most hyper-parameters for the five acoustic models are the same. The optimizer is Adam and dropout rate is 0.2. Initial learning rate is set to 10^{-3} for all acoustic models.

2) CHiME-2: For experiments on CHiME-2, the noise-mismatched acoustic model uses the well-trained noise-dependent acoustic model. Therefore, there are only four acoustic models on this corpus. The optimizer and dropout rate are the same as those on WSJ. The four acoustic models share the same initial learning rate of 10^{-4}.

IV. Evaluation Results and Analysis

This section presents and analyzes our evaluation results on the five acoustic models. The results are provided separately for the WSJ and CHiME-2 corpora.

A. Results on WSJ

TABLE I shows the WERs of the five acoustic models on WSJ. We use ADTbabble and ADTcafeteria1 as the noises for evaluation. The clean acoustic model clearly benefits from speech enhancement, as enhanced speech has a higher SNR than the corresponding noisy speech.

For the noise-dependent acoustic model, consistent with previous observations [4], [8], [14], the results on unenhanced speech are better. Based on our analysis on the distortion

TABLE I
WERS OF THE FIVE ACOUSTIC MODELS ON WSJ. *bab* AND *caf* DENOTE ADTBABBLE AND ADTCAFETERIA1, RESPECTIVELY. *w/o* REFERS TO NOISY EVALUATION DATA WITHOUT SPEECH ENHANCEMENT (I.E. UNENHANCED SPEECH), AND *w/* EVALUATION DATA WITH ENHANCEMENT.

SNR	clean				noise-dependent				noise-mismatched				noise-independent				distortion-independent			
	bab		caf		bab		caf		bab		caf		bab		caf		bab		caf	
	w/o	w/	w/o	w/	w/o	w/	w/o	w/	w/o	w/	w/o	w/	w/o	w/	w/o	w/	w/o	w/	w/o	w/
9dB	11.92	**3.08**	12.83	**3.53**	**3.62**	4.35	**4.28**	5.04	6.31	**5.01**	4.95	**4.05**	4.89	**4.00**	4.97	**4.04**	4.18	**3.10**	3.81	**3.29**
6dB	22.19	**4.11**	22.06	**6.15**	**4.28**	5.04	**5.55**	6.39	9.83	**5.94**	7.98	**5.77**	7.14	**4.86**	7.17	**5.55**	5.10	**4.00**	5.59	**4.80**
3dB	38.26	**6.67**	38.88	**9.15**	**5.12**	6.31	**8.11**	8.93	17.07	**7.85**	14.16	**8.59**	10.59	**6.65**	11.06	**8.09**	7.17	**5.23**	8.85	**7.08**
0dB	60.32	**12.46**	58.25	**17.34**	**7.55**	9.64	**12.07**	14.68	28.41	**11.68**	26.13	**14.05**	18.23	**10.74**	17.56	**14.18**	12.87	**9.19**	15.21	**12.85**
-3dB	82.44	**23.24**	79.15	**32.51**	**12.55**	18.03	**21.93**	27.72	46.16	**21.39**	48.48	**26.43**	31.89	**19.71**	31.14	**26.64**	24.30	**17.13**	27.82	**24.58**
-6dB	93.16	**44.76**	91.44	**56.25**	**22.66**	34.34	**40.13**	48.40	71.64	**38.05**	74.67	**49.52**	54.06	**36.19**	53.99	**47.94**	45.41	**33.55**	50.68	**45.17**
avg	51.4	**15.7**	50.4	**20.8**	**9.3**	13.0	**15.3**	18.5	29.9	**15.0**	29.4	**18.1**	21.1	**13.7**	21.0	**17.7**	16.5	**12.0**	19.4	**16.3**

problem, the performance degradation on enhanced speech is caused by the mismatch between N_{tr} and D_{eval}.

The noise-mismatched acoustic models are able to benefit from speech enhancement in our experiments on WSJ. Such an ability, however, is influenced by the type of noise used for testing. We will discuss this more after presenting the results on CHiME-2. In TABLE I, we observe that the results of noise-dependent acoustic models are much better than those of noise-mismatched acoustic models on unenhanced speech. This indicates that acoustic models trained on one noise cannot generalize to untrained noises. This performance degradation caused by noise mismatch supports our analysis on the distortion problem.

The noise-independent acoustic model also benefits from speech enhancement on WSJ. This indicates that 10k additive noises can capture a lot of the distortions on WSJ. The efficacy of noise-independent acoustic modeling, however, may be influenced by factors such as reverberation, as will be shown in the results on the CHiME-2 corpus.

For the distortion-independent acoustic model, the results on enhanced speech are better than those on unenhanced speech. This shows that distortion-independent acoustic models are able to alleviate the distortion problem caused by GRN. Moreover, the results of the distortion-independent acoustic model are better than those of the noise-independent model. Note that both noise-independent and distortion-independent acoustic models are tested on noises different from those used during training. The strong performance of our distortion-independent acoustic model shows that large-scale training with various distortions generalizes well to untrained distortions.

Along each column of TABLE I, there is a clear performance degradation as SNR reduces, consistent with our analysis on the cause of the distortion problem.

In TABLE II, we present the results of the distortion-independent acoustic model when coupled with speech enhancement frontends different from the one used during training. From the table, we observe that both LSTM and CRN yield better results than unenhanced speech. This shows that the distortion-independent acoustic model is able to generalize to different enhancement frontends. This also suggests that there may be a common pattern in the distortions introduced by supervised speech enhancement models.

Comparing the results of LSTM, CRN, and IRM, we find that for the distortion-independent acoustic model, the improvement of speech enhancement quality results in the improvement of recognition performance. In real-world applications, this suggests that a distortion-independent model need not to be retrained when a more advanced speech enhancement frontend is applied. In addition, the distortion-independent acoustic model on WSJ may be used to provide an indicator on the modeling ability of different speech enhancement frontends. Note that at different SNRs, the IRM results vary slightly, which may be due to the waveform resynthesis during speech enhancement. When the distortion-independent acoustic model is evaluated on clean speech, the average WER is 2.7%. For the clean acoustic model evaluated on clean speech, the WER is 2.0%.

B. Results on CHiME-2

TABLE III presents the WERs of the five acoustic models on CHiME-2. The noises used for evaluation include chime-2 noises and ADT. WERs on ADT are the averages of those on ADTbabble and ADTcafeteria1. The reverberant acoustic model on CHiME-2 corresponds to the clean acoustic model on WSJ. It is clear that the reverberant acoustic model benefits from speech enhancement.

The noise-dependent acoustic model follows the official training recipe of the CHiME-2 challenge. Similar to prior observations [4], [13], the noise-dependent acoustic model does not benefit from speech enhancement, which is in line with the results in TABLE I.

We test the noise-mismatched acoustic model on ADT noises. Different from the results on WSJ, the noise-mismatched acoustic model does not perform better on enhanced speech. Note that the experiments on CHiME-2 use CHiME-2 noises for training, whereas the experiments on WSJ

TABLE II
WERS OF THE DISTORTION-INDEPENDENT ACOUSTIC MODEL ON WSJ WITH OTHER FRONTENDS. SEE TABLE I CAPTION FOR NOTATIONS.

SNR	unenhanced		LSTM		CRN		IRM	
	bab	caf	bab	caf	bab	caf	bab	caf
9dB	4.18	3.81	3.36	3.21	3.27	4.09	2.73	2.73
6dB	5.10	5.59	4.24	4.88	4.13	4.65	2.75	2.88
3dB	7.17	8.85	5.70	7.64	5.03	7.53	2.84	2.65
0dB	12.87	15.21	9.02	13.17	8.39	11.53	2.88	2.86
-3dB	24.30	27.82	17.95	25.14	14.37	22.38	3.05	2.76
-6dB	45.41	50.68	34.99	47.80	28.19	40.82	2.91	2.93
avg	16.5	19.4	12.5	17.0	10.6	15.2	2.9	2.8

TABLE III
WERS OF THE FIVE ACOUSTIC MODELS ON CHIME-2. *chime-2* DENOTES THE OFFICIAL CHIME-2 EVALUATION SET. *ADT* REFERS TO THE AVERAGE WER OF ADTBABBLE AND ADTCAFETERIA1. SEE TABLE I CAPTION FOR OTHER NOTATIONS.

SNR	reverberant				noise-dependent		noise-mismatched		noise-independent				distortion-independent			
	chime-2		ADT		chime-2		ADT		chime-2		ADT		chime-2		ADT	
	w/o	w/	w/o	w/	w/o	w/	w/o	w/	w/o	w/	w/o	w/	w/o	w/	w/o	w/
9dB	31.27	**10.50**	31.03	**11.40**	**5.49**	5.81	**7.82**	8.33	6.63	**6.37**	**6.59**	7.98	7.42	**5.51**	10.20	**6.60**
6dB	38.69	**13.67**	47.53	**19.00**	**6.26**	7.98	**10.28**	11.76	**7.72**	7.92	**8.66**	11.04	8.61	**6.54**	13.27	**8.64**
3dB	46.85	**17.26**	67.50	**31.68**	**6.78**	8.33	**18.03**	20.46	8.82	**8.78**	**14.00**	19.35	10.01	**7.10**	20.99	**14.73**
0dB	57.33	**23.73**	85.96	**50.89**	**8.95**	11.26	**30.07**	34.38	**10.69**	11.62	**23.72**	30.99	12.93	**9.70**	32.30	**22.76**
-3dB	62.94	**29.91**	93.49	**71.89**	**9.98**	14.48	**50.34**	55.30	**13.06**	13.30	**39.04**	51.60	14.85	**11.04**	51.02	**37.74**
-6dB	72.31	**39.87**	95.58	**88.79**	**14.83**	19.05	**75.65**	78.83	**17.45**	19.80	**60.73**	76.08	21.80	**15.45**	75.54	**58.24**
avg	51.6	**22.5**	70.2	**45.6**	**8.7**	11.2	**32.0**	34.8	**10.7**	11.3	**25.5**	32.8	12.6	**9.2**	33.9	**24.8**

use ADT noises. The inconsistent results on the two corpora indicate that the ability of the noise-mismatched acoustic model to overcome the distortion problem may depend on the noise used for testing.

The noise-independent acoustic model on CHiME-2 does not gain performance improvement on enhanced speech. This is again different from the corresponding results on WSJ. On the CHiME-2 corpus, room impulse responses (RIRs) are different between training and testing [20]. Although we use a large variety of additive noises to train the noise-independent acoustic model, the RIR mismatch still exists. During testing, distortions introduced by speech enhancement thus deviate from the 10k additive noises used for training. Note that at SNR level 9dB and 3dB, enhanced speech performs better than unenhanced speech on the CHiME-2 corpus.

The distortion-independent acoustic model is able to benefit from speech enhancement. On both CHiME-2 and ADT noises, distortion-independent acoustic model outperforms noise-independent acoustic model. The ability of distortion-independent acoustic modeling to benefit from speech enhancement shows that large-scale training on a variety of distortions generalizes to untrained distortions.

TABLE IV shows the results of the distortion-independent acoustic model when used with different speech enhancement frontends. Similar to the experiments on WSJ, distortion-independent acoustic model is tested on LSTM and CRN enhanced speech. The results of both LSTM and CRN enhanced speech are better than those of unenhanced speech. This indicates the ability of the distortion-independent acoustic model to work with various speech enhancement frontends. This also suggests that distortions introduced by different supervised enhancement models have certain similarities.

On IRM enhanced speech, the distortion-independent acoustic model performs very well. This suggests that as speech enhancement research progresses, speech recognition performances of the distortion-independent acoustic model should also improve. The average WER of the distortion-independent acoustic model on reverberant speech is 3.4%. For the reverberant acoustic model evaluated on reverberant speech, the WER is 2.8%.

TABLE V shows a comparison of ASR systems in this study with those in prior work. It is worth noting that our distortion-independent acoustic model achieves a 9.2% WER, which is better than the previous best systems on the CHiME-2 corpus [16], [27]. For the noise-dependent acoustic model, we achieve an average WER of 8.7%, outperforming the previous

TABLE IV
WERs OF THE DISTORTION-INDEPENDENT ACOUSTIC MODEL ON CHiME-2 WITH OTHER FRONTENDS. SEE TABLE III CAPTION FOR NOTATIONS.

SNR	unenhanced		LSTM		CRN		IRM	
	chime-2	ADT	chime-2	ADT	chime-2	ADT	chime-2	ADT
9dB	7.42	10.20	5.79	7.61	6.65	7.50	3.40	3.66
6dB	8.61	13.27	7.47	10.21	7.68	10.09	3.44	3.64
3dB	10.01	20.99	8.63	17.57	9.04	15.57	3.34	3.62
0dB	12.93	32.30	11.36	28.47	11.25	25.73	3.38	3.73
-3dB	14.85	51.02	14.16	44.83	13.51	41.12	3.74	3.95
-6dB	21.80	75.54	19.41	67.08	18.06	62.33	3.31	4.10
avg	12.6	33.9	11.1	29.3	11.0	27.1	3.4	3.8

TABLE V
WER COMPARISONS BETWEEN THE PROPOSED MODELS AND PRIOR WORK ON CHiME-2.

models	9dB	6dB	3dB	0dB	-3dB	-6dB	avg
Wang and Wang [27]	6.61	6.86	8.67	10.39	13.02	18.23	10.6
Plantinga *et al.* [16]	-	-	-	-	-	-	9.3
distortion-independent	5.51	6.54	7.10	9.70	11.04	15.45	9.2
noise-dependent	**5.49**	**6.26**	**6.78**	**8.95**	**9.98**	**14.83**	**8.7**

best system by 6.5% relatively. Note that, in order to avoid the influence of model adaptation on our analysis of the distortion problem, we do not apply speaker adaptation to our models. The good results of our proposed models suggest that the observations in this study are likely valid for real world systems.

V. CONCLUDING REMARKS

The distortion problem occurs when we apply speech enhancement as a frontend for ASR tasks. This study treats the distortion problem as a noise mismatch between training and testing. We categorize acoustic models into five types and examine each of them for their ability to overcome the distortion problem. Distortion-independent acoustic modeling emerges as the best among the five acoustic models. Its ability to generalize to untrained noises suggests the utility of large-scale training for acoustic modeling. We also show that the distortion-independent acoustic model is able to work with various speech enhancement frontends. In addition, the WERs of our proposed distortion-independent and noise-dependent acoustic models both outperform the previous best system on the CHiME-2 corpus.

Future work on the distortion problem includes using ASR features as the training target of speech enhancement models, applying time-domain speech enhancement frontends, and investigating distortion-independent training for end-to-end ASR systems.

REFERENCES

[1] D. Bagchi, P. Plantinga, A. Stiff, and E. Fosler-Lussier, "Spectral feature mapping with mimic loss for robust speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2018 IEEE International Conference on*. IEEE, 2018, pp. 5609–5613.

[2] J. Chen and D. Wang, "Long short-term memory for speaker generalization in supervised speech separation," *The Journal of the Acoustical Society of America*, vol. 141, pp. 4705–4714, 2017.

[3] J. Chen, Y. Wang, S. Yoho, D. Wang, and E. Healy, "Large-scale training to increase speech intelligibility for hearing-impaired listeners in novel noises," *The Journal of the Acoustical Society of America*, vol. 139, pp. 2604–2612, 2016.

[4] J. Du, Q. Wang, T. Gao, Y. Xu, L. Dai, and C. Lee, "Robust speech recognition with speech enhanced deep neural networks," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014, pp. 616–620.

[5] T. Gao, J. Du, L. Dai, and C. Lee, "SNR-based progressive learning of deep neural network for speech enhancement." in *Proc. of INTERSPEECH*, 2016, pp. 3713–3717.

[6] T. Gao, J. Du, Y. Xu, L. Liu, C.and Dai, and C. Lee, "Improving deep neural network based speech enhancement in low SNR environments," in *International Conference on Latent Variable Analysis and Signal Separation*. Springer, 2015, pp. 75–82.

[7] E. Healy, S. Yoho, Y. Wang, and D. Wang, "An algorithm to improve speech recognition in noise for hearing-impaired listeners," *The Journal of the Acoustical Society of America*, vol. 134, pp. 3029–3038, 2013.

[8] J. Heymann, L. Drude, and H. Reinhold, "Wide residual BLSTM network with discriminative speaker adaptation for robust speech recognition," in *Proceedings of the 4th International Workshop on Speech Processing in Everyday Environments (CHiME16)*, 2016, pp. 12–17.

[9] M. Kolbk, Z. Tan, and J. Jensen, "Speech intelligibility potential of general and specialized deep neural network based speech enhancement systems," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 25, pp. 153–167, 2017.

[10] K. Li, Z. Huang, Y. Xu, and C. Lee, "DNN-based speech bandwidth expansion and its application to adding high-frequency missing features for automatic speech recognition of narrowband speech," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015, pp. 2578–2582.

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

[12] A. Narayanan, A. Misra, K. Sim, G. Pundak, A. Tripathi, M. Elfeky, P. Haghani, T. Strohman, and M. Bacchiani, "Toward domain-invariant speech recognition via large scale training," *arXiv preprint arXiv:1808.05312*, 2018.

[13] A. Narayanan and D. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 7092–7096.

[14] ——, "Investigation of speech separation as a front-end for noise robust speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, pp. 826–835, 2014.

[15] S. Nie, H. Zhang, X. Zhang, and W. Liu, "Deep stacking networks with time series for speech separation," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 6667–6671.

[16] P. Plantinga, D. Bagchi, and E. Fosler-Lussier, "An exploration of mimic architectures for residual network based spectral mapping," *arXiv preprint arXiv:1809.09756*, 2018.

[17] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The Kaldi speech recognition toolkit," in *Proc. of IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, no. EPFL-CONF-192584, 2011.

[18] K. Tan, J. Chen, and D. Wang, "Gated residual networks with dilated convolutions for monaural speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, pp. 189–198, 2019.

[19] K. Tan and D. Wang, "A convolutional recurrent neural network for real-time speech enhancement," in *Proc. of INTERSPEECH*, 2018, pp. 3229–3233.

[20] E. Vincent, J. Barker, S. Watanabe, J. Le Roux, F. Nesta, and M. Matassoni, "The second chime speech separation and recognition challenge: Datasets, tasks and baselines," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 126–130.

[21] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, pp. 1702–1726, 2018.

[22] P. Wang and D. Wang, "Filter-and-convolve: A CNN based multichannel complex concatenation acoustic model," in *Acoustics, Speech and Signal Processing (ICASSP), 2018 IEEE International Conference on*, 2018, pp. 5564–5568.

[23] ——, "Utterance-wise recurrent dropout and iterative speaker adaptation for robust monaural speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2018 IEEE International Conference on*, 2018, pp. 4814–4818.

[24] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 22, pp. 1849–1858, 2014.

[25] Y. Wang and D. Wang, "Boosting classification based speech separation using temporal dynamics," in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012, pp. 1528–1531.

[26] ——, "Cocktail party processing via structured prediction," in *Advances in Neural Information Processing Systems*, 2012, pp. 224–232.

[27] Z. Wang and D. Wang, "A joint training framework for robust automatic speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, pp. 796–806, 2016.

[28] F. Weninger, H. Erdogan, S. Watanabe, E. Vincent, J. Le Roux, J. 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.

[29] Y. Xu, J. Du, L. Dai, and C. Lee, "Dynamic noise aware training for speech enhancement based on deep neural networks," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014, pp. 2670–2674.

[30] ——, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal processing letters*, vol. 21, pp. 65–68, 2014.

[31] ——, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 23, pp. 7–19, 2015.

[32] Y. Xu, J. Du, Z. Huang, L. Dai, and C. Lee, "Multi-objective learning and mask-based post-processing for deep neural network based speech enhancement," *arXiv preprint arXiv:1703.07172*, 2017.