Speech-enhanced and Noise-aware Networks for Robust Speech Recognition

Hung-Shin Lee[1], *Pin-Yuan Chen*[1], *Yao-Fei Cheng*[1], *Yu Tsao*[2], *and Hsin-Min Wang*[1]

[1]Institute of Information Science, Academia Sinica
[2]Research Center for Information Technology Innovation, Academia Sinica

hungshinlee@gmail.com, freddy@iis.sinica.edu.tw

Abstract

Compensation for channel mismatch and noise interference is essential for robust automatic speech recognition. Enhanced speech has been introduced into the multi-condition training of acoustic models to improve their generalization ability. In this paper, a noise-aware training framework based on two cascaded neural structures is proposed to jointly optimize speech enhancement and speech recognition. The feature enhancement module is composed of a multi-task autoencoder, where noisy speech is decomposed into clean speech and noise. By concatenating its enhanced, noise-aware, and noisy features for each frame, the acoustic-modeling module maps each feature-augmented frame into a triphone state by optimizing the lattice-free maximum mutual information and cross entropy between the predicted and actual state sequences. On top of the factorized time delay neural network (TDNN-F) and its convolutional variant (CNN-TDNNF), both with SpecAug, the two proposed systems achieve word error rate (WER) of 3.90% and 3.55%, respectively, on the Aurora-4 task. Compared with the best existing systems that use bigram and trigram language models for decoding, the proposed CNN-TDNNF-based system achieves a relative WER reduction of 15.20% and 33.53%, respectively. In addition, the proposed CNN-TDNNF-based system also outperforms the baseline CNN-TDNNF system on the AMI task.

Index Terms: robust speech recognition, autoencoder, multi-condition training, noise-aware training

1. Introduction

For nearly a decade, the development of deep learning-based methods for automatic speech recognition (ASR) has advanced significantly, regardless of whether the acoustic model (AM) is trained using a deep neural network/hidden Markov model (DNN/HMM)-based process or a phone-free end-to-end (E2E) structure. In realistic situations, the speech input to an ASR system may include a considerable amount of interference from various types of background noise, and there may be a channel mismatch between it and the training speech. Therefore, it is very important to improve the robustness of ASR. This study is focused on boosting the noise robustness of Gaussian mixture model (GMM)/DNN-based ASR.

As described in [1], there are three deep learning-based approaches that can boost the robustness of ASR. The first approach uses speech enhancement (SE) models with various neural structures, such as the denoising autoencoder (DAE) [2, 3], dense neural network [4], and generative adversarial network (GAN) [5, 6], to pre-process the input speech. These SE models are usually trained to estimate the ideal ratio mask (IRM) for speech or directly restore clean speech from noisy speech. However, the training objective is not directly related to the ultimate goal of ASR, which is to minimize phone or word errors. Therefore, the robustness of the downstream ASR system largely depends on the performance of SE components [7, 8].

The second approach is multi-condition training (MCT), which uses clean and noisy speech for acoustic modeling. The purpose of MCT is to optimize the familiarity of the ASR system with various types of noisy speech to avoid mismatches between training and testing environments. Although MCT still suffers from unexpected conditions and speech distortion [9, 10], in several typical ASR competitions, such as the CHiME challenge, it has been widely used through data augmentation with noise addition or reverberation simulation [11]. In addition to noisy speech, the enhanced speech generated by SE models can be added to the training set [3]. As a special case of MCT, SpecAug directly applies certain spectrogram deformations to speech signals for data augmentation, which significantly improves the ASR performance [12].

The third approach is joint training (JT), in which an SE model consisting of an autoencoder [13, 14] or GAN [15] is attached to the AM to minimize the losses of SE and ASR in a unified manner. The ASR-related loss is mainly associated with connectionist temporal classification or attention-based sequence-to-sequence models in E2E systems [16, 17, 1]. As for the GMM/DNN topology, only the phoneme/state-level cross entropy (CE) has been adopted in the literature so far [13, 18, 8, 14]. JT compensates for the shortcomings of the first approach in several ways. By keeping the contributions of recognition and enhancement in balance, it prevents the enhanced speech from being over-smoothed. Enhanced speech itself is a good source of feature-level data augmentation. For example, Wang and Wang concatenated noisy mel-frequency cepstral coefficient (MFCC) features and enhanced filter-bank features as the input of the back-end AM [8], whereas Fan *et al.* proposed a gated recurrent fusion method that fuses noisy and enhanced features to address speech distortion [1]. Moreover, in [13, 14, 16, 17], only the enhanced features were used as the input of AM. In this case, the SE loss functions as a sample-level regularizer to avoid overfitting.

In addition to the above-mentioned methods with the help of SE, noise-aware training (NAT), which appends "noise vectors" to the input feature sequence, has also attracted attention in robust ASR [9, 19, 20]. The main difference between them is whether the noise vectors are estimated through noise detection [9, 20] or directly derived from noisy speech [19]. Noise representations in NAT is learnable using neural networks. NAT can also be thought of as a type of feature-level data augmentation.

In this paper, we propose a unified framework that combines MCT, JT, and NAT to achieve robust ASR. The framework is implemented on GMM/DNN AMs trained using the lattice-free maximum mutual information (LF-MMI) criterion [21]. As shown in Table 1, the main novelty of this study is twofold. First, the simultaneous optimization of the frame-level cross-entropy, utterance-level LF-MMI, and acoustic feature reconstruction error is unprecedented in JT-based robust ASR.

Table 1: *JT-based methods for robust ASR. EnhC and Enh represent the enhancement component and the enhanced feature, respectively. G/D denotes the GMM/DNN topology.*

Method	Topo.	EnhC	Augment.
Gao *et al.* (2015) [13]	G/D	DAE	-
Mimura *et al.* (2016) [14]	G/D	DAE	-
Wang *et al.* (2016) [8]	G/D	T/F IRM	Enh
Kundu *et al.* (2016) [19]	G/D	FL	Enh & Noise
Liu *et al.* (2018) [15]	E2E	GAN	-
Liu *et al.* (2019) [16]	E2E	T/F PSA	-
Soni *et al.* (2019) [17]	G/D	T/F IRM	-
Fan *et al.* (2021) [1]	E2E	T/F IAM	Enh
Ours	G/D	DAE	Enh & Noise

Second, the introduction of noise-aware training and data augmentation with noise not only strengthens the capability of the enhancement component to extract purer enhanced speech, but also enables the subsequent AM to utilize the estimated noise feature to filter out the noise in noisy speech more effectively.

2. Proposed Models

Fig. 1 illustrates the schematic diagram of our proposed neural architecture, where two additional components and AM are sequentially connected and trained jointly. This work has been open-sourced at https://github.com/Sinica-SLAM/kaldi-SENAN.

2.1. Acoustic Model

The AM converts a frame representation $\mathbf{x}_{in}$ into triphone-state scores. The input $\mathbf{x}_{in}$ is the concatenation of three types of features: noisy features and the i-vector $\mathbf{x}_{nsy}$, enhanced features $\mathbf{x}_{enh}$, and noise-aware features $\mathbf{x}_{nse}$.

$$\mathbf{x}_{in} = \mathbf{x}_{nsy} \oplus \mathbf{x}_{enh} \oplus \mathbf{x}_{nse}, \tag{1}$$

where $\oplus$ denotes the concatenation operator.

The AM, parameterized by θ, generates the triphone-state scores (see the top black block in Fig. 1). In this study, the AM was implemented on a factorized time-delay neural network (TDNN-F) or its variant with convolutional layers (CNN-TDNNF) [22, 23, 24]. We derived two ASR-related objective functions. The first is the CE between the predicted triphone-state scores and ground truth defined by

$$\mathcal{L}_{CE} = -\sum_{u=1}^{U}\sum_{t=1}^{T_u} \log p_\theta(\mathbf{s}^{ut}|\mathbf{x}_{in}^{ut}), \tag{2}$$

where U is the number of training utterances, T_u is the number of frames of utterance u, and $\mathbf{s}^{ut}$ and x_{in}^{ut} are the true triphone-state label and acoustic input of the t-th frame of utterance u, respectively. The label is derived by GMM-based forced alignment. The second is based on MMI, a discriminative criterion designed to maximize the probability of the reference transcription while minimizing the probabilities of all other transcriptions [25]. As in [26], the LF-MMI is defined as

$$\mathcal{F}_{LF-MMI} = \sum_{u=1}^{U} \log \frac{p_\theta(\mathbf{x}_{in}^u|M_{\mathbf{w}_u})P(M_{\mathbf{w}_u})}{p_\theta(\mathbf{x}_{in}^u|M_{den})}, \tag{3}$$

where $\mathbf{w}_u$ is the reference word sequence of utterance u. The composite HMM graph $M_{\mathbf{w}_u}$ represents all possible state sequences pertaining to $\mathbf{w}_u$ and is called the numerator graph.

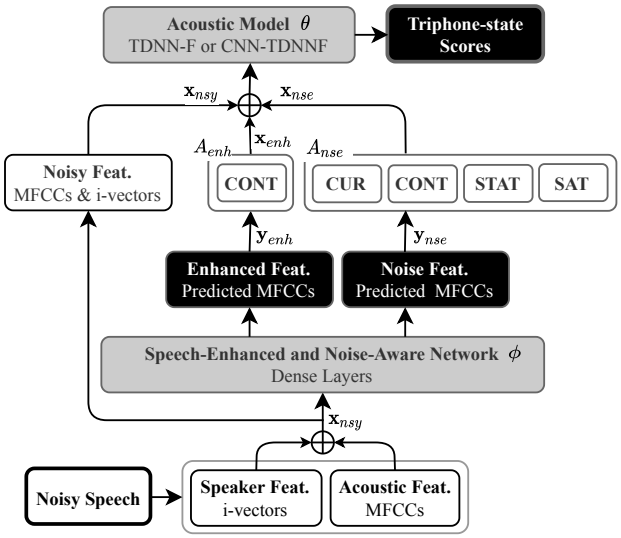


Figure 1: *The proposed speech-enhanced and noise-aware acoustic model, where the three black blocks are directly related to loss functions during training.*

M_{den} is the denominator graph that models all the possible sequences. It has traditionally been estimated using lattices because the full denominator graph can become large and significantly slow the computation. Povey *et al.* derived MMI training of DNN/HMM models using a full denominator graph (hence, the name lattice-free) [21].

2.2. Speech-enhanced and Noise-aware Network (SENAN)

If $\mathbf{x}_{enh}$ and $\mathbf{x}_{nse}$ in Eq. 1 are discarded, our proposed model is reduced to the conventional MCT-based framework. The speech-enhanced and noise-aware network (SENAN) ϕ, constructed with dense layers, decomposes the frame-wise noisy features $\mathbf{x}_{nsy}$ into the speech-enhanced and noise-aware parts, i.e., $\mathbf{y}_{enh}$ and $\mathbf{y}_{nse}$. Then, the aggregation functions output $\mathbf{x}_{enh}$ and $\mathbf{x}_{nse}$ from $\mathbf{y}_{enh}$ and $\mathbf{y}_{nse}$ by

$$[\mathbf{x}_{enh}, \mathbf{x}_{nse}] = [A_{enh}(\mathbf{y}_{enh}), A_{nse}(\mathbf{y}_{nse})], \tag{4}$$

where $[\mathbf{y}_{enh}, \mathbf{y}_{nse}] = \phi(\mathbf{x}_{nsy})$, and $A(\cdot)$ denotes the aggregation function, which is used to leverage the contextual or statistical information near the working frame. As shown in Fig. 1, the aggregation function can consider the current frame only (CUR), the current frame appended with one preceding and one following frame (CONT), the concatenation of two statistical measurements (the mean and variance vectors) derived from a context window of 150 frames (STAT) [27], and single-head self-attention in a local range of five preceding and two following frames around the current frame (SAT) [28]. Based on preliminary results, we use CONT as the aggregation function for $\mathbf{y}_{enh}$, and evaluate the four functions for $\mathbf{y}_{nse}$ in this paper.

To ensure that the enhanced features $\mathbf{y}_{enh}$ and estimated noise $\mathbf{y}_{nse}$ are as similar as possible to the corresponding clean speech and true noise, respectively, we minimize the mean squared error (MSE) defined by

$$\mathcal{L}_{MSE}^{\{enh,nse\}} = \sum_{u=1}^{U}\sum_{t=1}^{T_u} \left|\left|\mathbf{y}_{\{enh,nse\}}^{ut} - \hat{\mathbf{y}}_{\{clean,nse\}}^{ut}\right|\right|_2^2, \tag{5}$$

where $||\cdot||_2^2$ is the 2-norm operator. $\mathbf{y}_{enh}^{ut}$ is the predicted enhanced MFCC vector, and $\hat{\mathbf{y}}_{clean}^{ut}$ is the corresponding MFCC

vector extracted from clean speech. $\mathbf{y}_{nse}^{ut}$ is the predicted noise MFCC vector, and $\hat{\mathbf{y}}_{nse}^{ut}$ is the corresponding MFCC vector extracted from the true noise signal.

2.3. Final Objective Function

The final objective function L to be minimized by backpropagation is a combination of ASR-related, speech-enhanced, and noise-aware losses, and is expressed as

$$\mathcal{L} = \alpha\mathcal{L}_{CE} - \mathcal{F}_{LF-MMI} + \beta(\mathcal{L}_{MSE}^{enh} + \mathcal{L}_{MSE}^{nse}), \quad (6)$$

where α and β are weighting factors. In previous studies in [29, 30, 31], only the CE and MSE metrics were jointly optimized in AM training. In our experiments, we set α to 5 as in most previous studies and heuristically set β to 0.2.

3. Experimental Settings

We used the Kaldi toolkit [32] to implement our models, and evaluated them on two datasets: the Aurora-4 corpus [33] and the AMI meeting corpus [34].

3.1. Datasets and Derivation of Noise

Aurora-4 [33] is a medium-level vocabulary task. The transcriptions were based on the Wall Street Journal corpus (WSJ0) [35]. The dataset contains 16 kHz speech data with additive noise and linear convolutional channel distortion, which were synthetically introduced into clean speech. The training set contains 7,138 utterance pairs from 83 speakers, where each pair consists of clean speech and speech corrupted by one of six different noises at 10–20 dB SNR. The test set was generated by the same types of noise and microphones and was grouped into four subsets: clean, noisy, clean with channel distortion, and noisy with channel distortion, which are referred to as A, B, C, and D, respectively. The speech data were recorded by two microphones, one of which was a near-field microphone. The speech recorded by this microphone can be considered clean and noise-free. In contrast, the speech recorded by the second microphone can be considered noisy speech. According to the assumption that noisy speech can be represented by the sum of a speech signal and a noise signal [36], we adjusted the volume of the noisy speech and subtracted the clean speech from the noisy speech to obtain the corresponding noise as the ground truth.

The AMI task is more challenging than the Aurora-4 task. The AMI corpus [34] contains approximately 77 h of meeting recordings, including 107 k utterances for training and 12.6 k utterances (8.7 h) for testing. The speech signals were captured and synchronized by multiple microphones, including individual head microphones, lapel microphones, and one or more microphone arrays. In this task, we took the speech recorded by individual head microphones and the standard SDM (single distant mic.) as the clean speech and the corresponding noisy speech, respectively. The method used for obtaining the noise was the same as that used for the Aurora-4 task.

3.2. Data Preparation and Preprocessing

Our training data (noisy and clean speech) were first organized for standard GMM training/alignment to derive the frame-wise triphone-state ground truth. To avoid overfitting the neural network to improve its robustness, speed and volume perturbations [37] were used to triple the amount of the training data. The trigram language model (LM) provided by the Kaldi recipe with a closed 5 k vocabulary was used in the Aurora-4 task, whereas the trigram LM trained with the transcriptions of the training set was used in the AMI task.

To extract acoustic features, spectral analysis was applied to 25-ms speech frames every 10 ms. For each frame, 40 high-resolution MFCCs were derived by discrete cosine transform conducted on 40 mel-frequency bins. Utterance-level mean subtraction was used for feature normalization. The noisy features were formed by concatenating the acoustic features and the 100-dimensional speaker-related i-vector [38].

Table 2: *WERs (%) on Aurora-4 with respect to different aggregation functions for the predicted noise features. AM: TDNN-F without SpecAug.*

Aggregation Func.	A	B	C	D	Avg.
CURR	1.61	3.59	3.12	9.88	6.11
CONT	1.87	3.53	3.34	9.61	6.00
STAT	1.74	3.29	3.01	9.60	**5.86**
ATTN	1.77	3.41	3.03	9.53	5.89

3.3. Neural Structures

3.3.1. Speech-enhanced and noise-aware network

We used the multi-task autoencoder (MTAE) in [39] to construct the SENAN ϕ in Fig. 1. MTAE has triangular-shaped shared units, and is a unified structure with a denoising autoencoder and a "de-speeching" autoencoder. There are 1,024 and 2,048 nodes in the first and last hidden layers, respectively. The number of nodes increases linearly from low to high hidden layers, and the number of hidden layers is five.

3.3.2. Acoustic model

One of the best neural structures for realizing the AM θ in Fig. 1 is the factorized form of TDNN [23], called TDNN-F. TDNN-F and TDNN have the same structure, but the layers of the former are compressed by singular value decomposition and trained by random initialization with semi-orthogonal constraints in one of the two factors of each matrix [24]. Our TDNN-F was constructed with 13 hidden layers, each containing 1,024 hidden nodes and 128 linear bottleneck nodes. The parameters of the final layer were factorized using a 192-node linear bottleneck layer. The mini-batch sizes were set to 128 and 64. The initial and final effective learning rates were set to 0.01 and 0.001, respectively, and the total number of training epochs was set to 20. In our CNN-TDNNF model, there were six convolutional layers followed by nine TDNN-F layers. The time offsets and height offsets were set to "-1, 0, 1" in terms of Kaldi for all convolutional layers, and the number of filters in the six convolutional layers was set to 48, 48, 64, 64, 64, and 128, respectively. The parameters of the TDNN-F layers and hyperparameters were set to the same value as in the TDNN-F system. The total number of training epochs was set to 10.

In addition, we applied SpecAug to the input of the AM during training. It involves deformation in two dimensions: temporal and feature-level masking.

4. Experimental Results

4.1. Primary Results

First, we compared four types of aggregation functions for the predicted noise features in Fig. 1 on Aurora-4. Table 2 shows that STAT outperforms the other three functions, presumably

Table 3: *WERs (%) on Aurora-4 with respect to TDNN-F and CNN-TDNNF. -/+ denote without/with, respectively. Proposed† is the oracle case, where SENAN is removed and the enhanced/noise features (i.e., $\mathbf{y}_{enh}$/$\mathbf{y}_{nse}$ in Eq. 4) are replaced by $\hat{\mathbf{y}}_{clean}$/$\hat{\mathbf{y}}_{nse}$ derived from clean speech and true noise.*

TDNN-F		**A**	**B**	**C**	**D**	**Avg.**
-SpecAug	Baseline	1.74	3.83	3.29	10.31	6.42
	Proposed	1.74	3.29	3.01	9.60	**5.86**
	Proposed†	1.55	1.60	1.63	1.62	1.61
+SpecAug	Baseline	1.33	2.55	2.15	6.28	4.03
	Proposed	1.59	2.45	2.30	6.01	**3.90**
CNN-TDNNF		**A**	**B**	**C**	**D**	**Avg.**
-SpecAug	Baseline	1.55	3.28	3.03	9.49	**5.80**
	Proposed	1.85	3.57	2.97	9.32	5.87
+SpecAug	Baseline	1.31	2.27	1.89	5.74	3.66
	Proposed	1.23	2.13	2.02	5.61	**3.55**

Table 4: *State-of-the-art benchmarks on Aurora-4 with respect to bigram and trigram language models used for decoding.*

	Method	**A**	**B**	**C**	**D**	**Avg.**
bi-gram	CAT-VDCRN [40]	2.95	4.17	3.70	8.55	5.92
	Proposed	2.20	3.42	3.34	7.36	**5.02**
tri-gram	CNN-Raw [41]	2.70	4.40	4.00	6.40	5.10
	Proposed	1.14	2.14	1.91	5.25	**3.39**

because statistical information can more consistently describe the noise pattern in noisy speech. Therefore, STAT will be used in the following experiments.

Then, we evaluated the WERs of different AMs on Aurora-4. Note that our proposed model can be regarded as a combination of the baseline and SENAN. Several observations can be made from Table 3. First, when using TDNN-F as AM, with or without SpecAug, our proposed model outperforms the baseline in most test subsets. The average WER is reduced by 8.72% (from 6.42% to 5.86% with SpecAug) and 3.23% (from 4.03% to 3.90% without SpecAug). Although the WER reduction is relatively small when SpecAug is used, the system combining SpecAug and SENAN still achieves the best performance (WER=3.90%). The results show that the two data augmentation techniques can complement each other. Second, when CNN-TDNNF is used as AM but without SpecAug, our proposed model performs slightly worse than the baseline because of its weak performance in Sets A and B. The results imply that our proposed model may underfit the training set, which was recorded using the same Sennheiser microphone as Sets A and B. However, when viewed optimistically, our proposed model is still better than the baseline in terms of handling real situations, as shown by the results of Sets C and D. Among all models, the proposed model combining CNN-TDNNF, SENAN, and SpecAug achieves the lowest WER of 3.55%. Third, in the oracle experiment, the average WER reached 1.61%, where the estimated clean speech and noise are replaced by ground truth. This result confirms that we are on the right track, but our model still has considerable room for improvement.

Table 5: *WERs (%) and relative changes over Baseline on AMI.*

Method	**Dev**	rel. %	**Eval**	rel. %
CNN-TDNNF (Baseline)	33.6	-	37.0	-
+ SpecAug	33.2	1.19	35.9	2.97
+ SENAN (Proposed)	**32.2**	4.17	**35.5**	4.05

Table 6: *Ablation studies on Aurora-4, where AGG is the aggregation function, and "+" means accumulating.*

Method	**WER %**	rel. %
TDNNF (Baseline)	6.42	-
+ SENAN (enhanced only)	6.17	3.89
+ SENAN (enhanced only w/ AGG)	6.08	5.30
+ SENAN (enhanced & noise w/ AGG)	5.86	8.72
+ SpecAug	3.90	39.25
+ Replaced with CNN-TDNNF	**3.55**	44.70

Next, we compared our best model (CNN-TDNNF+SENAN+SpecAug) with the best methods in the literature for the Aurora-4 task. Following the practice of these methods, the forced alignment of noisy speech before training neural AMs was directly obtained from that of the corresponding *clean* speech. The results in Table 4 confirm the superiority of the proposed model. Note that the subtle WER difference (3.55% in Table 3 vs. 3.39% in Table 4) is due to the different forced alignment of noisy training speech.

Finally, we evaluated the proposed model on the more challenging corpus AMI. The results in Table 5 again demonstrate that both SpecAug and SENAN can improve recognition performance, and they complement each other. It is also worth noting that standing on the shoulders of SpecAug, SENAN can further substantially reduce the WER on the Dev set ($4.17\%-1.19\% = \mathbf{2.98}\% > \mathbf{1.19}\% = 1.19\% - 0\%$).

4.2. Ablation Studies

Our proposed model contains two key components: SENAN and the aggregation function. SENAN is reduced to a denoising autoencoder if the predicted noise feature $\mathbf{y}_{nse}$ in Fig. 1 is discarded. We conducted a series of ablation studies by adding or changing network components to demonstrate their efficacy. The results in Table 6 show that each component does contribute to the performance.

5. Conclusion

In this paper, we have proposed an architecture that combines speech enhancement and acoustic modeling to achieve robust ASR and a joint training mechanism that involves ASR-related losses, in particular the lattice-free MMI, and feature reconstruction errors. The key component, SENAN, plays an important role in separately extracting enhanced speech and noise features from noisy speech. Through appropriate aggregation functions, the two features extracted by SENAN can be combined with the noisy speech feature to boost the ASR performance. They enable acoustic models to filter out non-speech content from noisy speech in a more effective way than conventional SE plus AM frameworks. Our experimental results have shown that the proposed models outperform the baseline models without SENAN on the Aurora-4 and AMI tasks, and outperform the state-of-the-art methods on the Aurora-4 task.

6. References

[1] C. Fan, J. Yi, J. Tao, Z. Tian, B. Liu, and Z. Wen, "Gated recurrent fusion with joint training framework for robust end-to-end speech recognition," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 198–209, 2021.

[2] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising," in *Proc. ICML*, 2008.

[3] M. Fujimoto and H. Kawai, "One-pass single-channel noisy speech recognition using a combination of noisy and enhanced features," in *Proc. Interspeech*, 2019.

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

[5] S. Pascual, A. Bonafonte, and J. Serra, "SEGAN: speech enhancement generative adversarial network," in *Proc. Interspeech*, 2017.

[6] C. Fan, B. Liu, J. Tao, J. Yi, Z. Wen, and Y. Bai, "Noise prior knowledge learning for speech enhancement via gated convolutional generative adversarial network," in *Proc. APSIPA ASC*, 2019.

[7] K. Han, Y. Wang, D. L. Wang, W. S. Woods, I. Merks, and T. Zhang, "Learning spectral mapping for speech dereverberation and denoising," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 6, pp. 982–992, 2015.

[8] Z.-Q. Wang and D. Wang, "A joint training framework for robust automatic speech recognition," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 4, pp. 796–806, 2016.

[9] M. L. Seltzer, D. Yu, and Y. Wang, "An investigation of deep neural networks for noise robust speech recognition," in *Proc. ICASSP*, 2013.

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

[11] V. Manohar, S.-j. Chen, Z. Wang, Y. Fujita, S. Watanabe, and S. Khudanpur, "Acoustic modeling for overlapping speech recognition: JHU CHiME-5 challenge system," in *Proc. ICASSP*, 2019.

[12] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A simple data augmentation method for automatic speech recognition," in *Proc. Interspeech*, 2019.

[13] T. Gao, J. Du, L.-R. Dai, and C.-H. Lee, "Joint training of front-end and back-end deep neural networks for robust speech recognition," in *Proc. ICASSP*, 2015.

[14] M. Mimura, S. Sakai, and T. Kawahara, "Joint optimization of denoising autoencoder and DNN acoustic model based on multi-target learning for noisy speech recognition," in *Proc. Interspeech*, 2016.

[15] B. Liu, S. Nie, Y. Zhang, D. Ke, S. Liang, and W. Liu, "Boosting noise robustness of acoustic model via deep adversarial training," in *Proc. ICASSP*, 2018.

[16] B. Liu, S. Nie, S. Liang, W. Liu, M. Yu, L. Chen, S. Peng, and C. Li, "Jointly adversarial enhancement training for robust end-to-end speech recognition," in *Proc. Interspeech*, 2019.

[17] M. Soni and A. Panda, "Label driven time-frequency masking for robust continuous speech recognition," in *Proc. Interspeeccch*, 2019.

[18] A. Narayanan and D. Wang, "Joint noise adaptive training for robust automatic speech recognition," in *Proc. ICASSP*, 2015.

[19] S. Kundu, G. Mantena, Y. Qian, T. Tan, M. Delcroix, K. C. Sim, and K. Yu, "Joint acoustic factor learning for robust deep neural network based automatic speech recognition," in *Proc. ICASSP*, 2016.

[20] D. Raj, J. Villalba, D. Povey, and S. Khudanpur, "Frustratingly easy Noise-aware training of acoustic models," 2020. [Online]. Available: http://arxiv.org/abs/2011.02090

[21] D. Povey, V. Peddinti, D. Galvez, P. Ghahremani, V. Manohar, X. Na, Y. Wang, and S. Khudanpur, "Purely sequence-trained neural networks for ASR based on lattice-free MMI," in *Proc. Interspeech*, 2016.

[22] V. Peddinti, G. Chen, V. Manohar, T. Ko, D. Povey, S. Khudanpur, and T. Johns, "JHU Aspire Syatem: Robust LVCSR with TDNNs, ivector adaptation and RNN-LMS," in *Proc. IEEE ASRU*, 2015.

[23] V. Peddinti, D. Povey, and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *Proc. Interspeech*, 2015.

[24] D. Povey, G. Cheng, Y. Wang, K. Li, H. Xu, M. Yarmohamadi, and S. Khudanpur, "Semi-orthogonal low-rank matrix factorization for deep neural networks," in *Proc. Interspeech*, 2018.

[25] K. Veselý, A. Ghoshal, L. Burget, and D. Povey, "Sequence-discriminative training of deep neural networks," in *Proc. Interspeech*, 2013.

[26] H. Hadian, H. Sameti, D. Povey, and S. Khudanpur, "Flat-start single-stage discriminatively trained HMM-based models for ASR," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 11, pp. 1949–1961, 2018.

[27] I. Medennikov, M. Korenevsky, T. Prisyach, Y. Khokhlov, M. Korenevskaya, I. Sorokin, T. Timofeeva, A. Mitrofanov, A. Andrusenko, I. Podluzhny, A. Laptev, and A. Romanenko, "The STC system for the CHiME-6 challenge," in *Proc. CHiME*, 2020.

[28] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. NeurIPS*, 2017.

[29] M.-H. Yang, H.-S. Lee, Y.-D. Lu, K.-Y. Chen, Y. Tsao, B. Chen, and H.-M. Wang, "Discriminative autoencoders for acoustic modeling," in *Proc. Interspeech*, 2017.

[30] P.-T. Huang, H.-S. Lee, S.-S. Wang, K.-Y. Chen, Y. Tsao, and H.-M. Wang, "Exploring the encoder layers of discriminative autoencoders for LVCSR," in *Proc. Interspeech*, 2019.

[31] H.-S. Lee, Y.-H. Peng, P.-T. Huang, Y.-C. Tseng, C.-H. Wu, Y. Tsao, and H.-M. Wang, "The Academia Sinica systems of speech recognition and speaker diarization for the CHiME-6 challenge," in *Proc. CHiME*, 2020.

[32] D. Povey, G. Boulianne, L. Burget, P. Motlicek, and P. Schwarz, "The Kaldi speech recognition toolkit," in *Proc. IEEE ASRU*, 2011.

[33] N. Parihar and J. Picone, "Aurora Working Group: DSR front end LVCSR evaluation AU/384/02," Inst. for Signal & Inform. Process., Mississippi State Univ., Tech. Rep., 2002. [Online]. Available: http://aurora.hsnr.de/aurora-4.html

[34] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, I. McCowan, W. Post, D. Reidsma, and P. Wellner, "The AMI meeting corpus: A pre-announcement," in *Proc. MLMI*, 2006.

[35] D. B. Paul and J. M. Baker, "The design for the Wall Street Journal-based CSR corpus," in *Proc. HLT*, 1992.

[36] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoustics, Speech, Signal Process.*, vol. 27, no. 2, pp. 113–120, 1979.

[37] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," in *Proc. Interspeech*, 2015.

[38] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 4, pp. 788–798, 2011.

[39] H. Zhang, C. Liu, N. Inoue, and K. Shinoda, "Multi-task autoencoder for noise-robust speech recognition," in *Proc. ICASSP*, 2018.

[40] Y. Qian, T. Tan, H. Hu, and Q. Liu, "Noise robust speech recognition on Aurora4 by humans and machines," in *Proc. ICASSP*, 2018.

[41] E. Loweimi, P. Bell, and S. Renals, "On the robustness and training dynamics of raw waveform models," in *Proc. Interspeech*, 2020.