Improving Deep Neural Network Based Speech Enhancement in Low SNR Environments

Tian Gao[1], Jun Du[1], Yong Xu[1], Cong Liu[2], Li-Rong Dai[1], Chin-Hui Lee[3⋆]

[1] University of Science and Technology of China, Hefei, Anhui, P. R. China
[2] iFlytek Research, iFlytek Co., Ltd., Hefei, Anhui, P. R. China
[3] Georgia Institute of Technology, Atlanta, Georgia, USA,
{gtian09,xuyong62}@mail.ustc.edu.cn, {jundu,lrdai}@ustc.edu.cn,
congliu2@iflytek.com, chl@ece.gatech.edu

Abstract. We propose a joint framework combining speech enhancement (SE) and voice activity detection (VAD) to increase the speech intelligibility in low signal-noise-ratio (SNR) environments. Deep Neural Networks (DNN) have recently been successfully adopted as a regression model in SE. Nonetheless, the performance in harsh environments is not always satisfactory because the noise energy is often dominating in certain speech segments causing speech distortion. Based on the analysis of SNR information at the frame level in the training set, our approach consists of two steps, namely: (1) a DNN-based VAD model is trained to generate frame-level speech/non-speech probabilities; and (2) the final enhanced speech features are obtained by a weighted sum of the estimated clean speech features processed by incorporating VAD information. Experimental results demonstrate that the proposed SE approach effectively improves short-time objective intelligibility (STOI) by 0.161 and perceptual evaluation of speech quality (PESQ) by 0.333 over the already-good SE baseline systems at -5dB SNR of babble noise.

Keywords: speech enhancement, low SNR, deep neural networks, voice activity detection, speech intelligibility

1 Introduction

Speech enhancement (SE) has been an open research problem for the past several decades. Many approaches are developed to solve this problem, and they can be classified into two categories, namely unsupervised and supervised methods. As for the unsupervised approaches, there are, spectral subtraction [1], MMSE-based log-spectral amplitude estimator [2] and optimally modified log-MMSE estimator [3], etc. However, many assumptions were made during the derivation process of these solutions, and the resulting enhanced speech often suffers from an annoying artifact called musical noise.

⋆ This work was supported by the National Natural Science Foundation of China under Grants No. 61305002. We would like to thank iFLYTEK Research for providing the training data and DNN training platform.

Various supervised methods have also been developed in recent years, which have been demonstrated to generate enhanced speech with better quality. Non-negative matrix factorization (NMF) based SE [4] was one of the notable methods. Speech and noise basis were learned from the speech data and noise data, respectively. Then the clean speech could be decomposed given the noisy speech. In [5, 6], masking techniques were used to train DNNs for speech separation and recognition. More recently, our proposed DNN-based SE where the DNN was regarded as a regression model to predict the clean log-power spectra (LPS) [7] from the noisy LPS has been successfully applied to noisy speech enhancement [8, 9], separation [10] and recognition [11, 12].

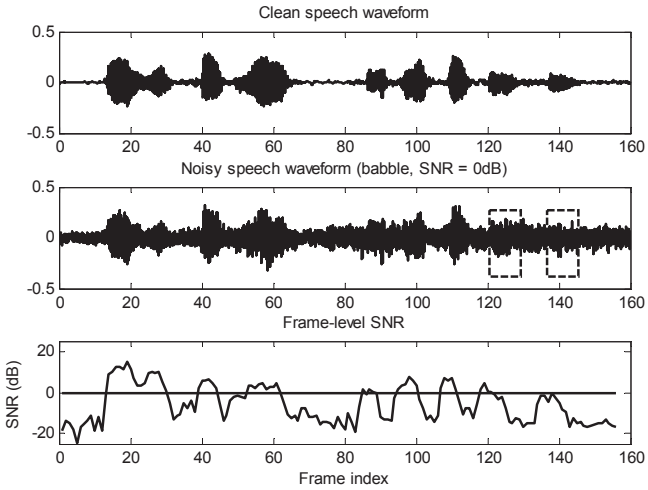

Fig. 1. *Illustration of an utterance example in the babble noise environment at SNR = 0dB along with the corresponding clean speech and frame-level SNR sequence.*

Fig. 1 shows noisy speech mixed with babble noise from the NOISEX-92 [13] corpus at SNR = 0dB along with the corresponding clean speech and frame-level SNR sequence. Speech segment covered by high-energy noises, such as the noted part in Fig. 1, remains difficult to handle. When noise is removed from those segments by conventional DNN approaches, the quality of speech is also severely degraded as it is not easy for a DNN to distinguish in those segments between speech and noise. The noisy speech segments with very weak speech energy are very similar to those pure noise segments in terms of frame-level SNR, which is a challenge for the data-driven approaches using a single DNN. In the frame-level DNN-based SE, local SNR distribution is more meaningful than global (e.g. utterance-level) for learning convergence. From the Fig. 1, we observe that the frame-level SNR values have a high fluctuation from the global SNR at 0dB. This indicates that the training set with a fixed, global SNR is multifarious at

frame level especially in low SNR conditions, and it will undoubtedly increase the difficulty of model learning.

In this paper, we propose a combined VAD+SE framework using DNNs in low SNR environments. The main contributions of this paper are summarized as follows: (i) We employ a system with dual outputs of speech features for both target and interference sources in the output layer as our baseline. (ii) We use the speech segments of the multi-condition training set using VAD [14, 15, 16] annotations from the corresponding clean speech to train a conservative speech enhancement (denoted as CSE) DNN model to well preserve the weak-energy speech segments in low SNR environments and conservatively remove the pure noise segments. (iii) A DNN-based VAD model is trained for system fusion. Empirical results demonstrate that the proposed framework can significantly improve the performance in low SNR environments.

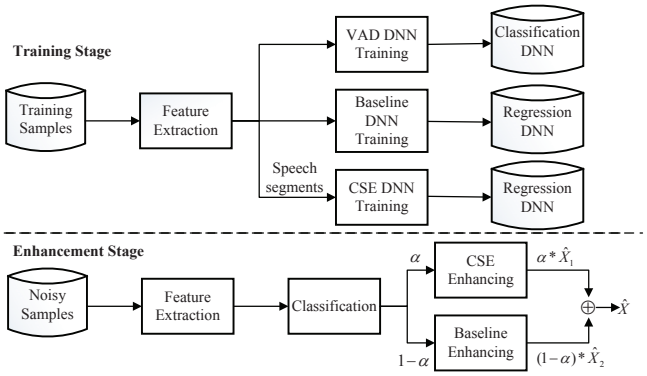

Fig. 2. *The proposed system.*

2 System Overview

The overall flowchart of the proposed SE system is illustrated in Fig. 2. First, the acoustic features of both clean speech and synthesized noisy speech training data are extracted. Then three DNNs, namely VAD DNN, baseline DNN and CSE DNN, are trained. In the enhancement stage, after feature extraction of the noisy utterance, frame-level soft decision is first given by the DNN-based VAD. To achieve better VAD performance, a long-term smoothing of the multiple DNN outputs with a half-window size τ can be applied. The classification DNN with smoothing is quite similar to the boosted DNN proposed in [17]. Then both the noisy features and speech/non-speech probabilities are presented to CSE and baseline system simultaneously. A fusion is performed with VAD classification probability to obtain the final enhanced speech signals as shown in Fig. 2. α is the probability of speech class, and $(1-\alpha)$ belong to the non-speech class. $\hat{\boldsymbol{X}}$,

$\hat{\boldsymbol{X}}_1$ and $\hat{\boldsymbol{X}}_2$ are the vectors of final enhanced speech, enhanced speech processed by CSE and by baseline system, respectively. This fusion can smooth the final enhanced speech and improve system performance. The details of both regression and classification DNNs are elaborated in Section 3.

3 DNN-based VAD and Speech Enhancement

3.1 DNN-based VAD

DNN for VAD is designed as a classification model where the output refers to the probabilities of two classes. The input to DNN is the noisy LPS features with neighboring frames. The training of this DNN consists of unsupervised pre-training and supervised fine-tuning. The former treats each consecutive pair of layers as a restricted Boltzmann machine (RBM) while the parameters of RBM are trained layer by layer with the approximate contrastive divergence algorithm [18]. After pre-training for initializing the weights of the first several layers, supervised fine-tuning of the parameters in the whole network is performed via a frame-level cross-entropy criterion. The main difference from other DNN approaches, e.g. [17], is the training data. In [17], only three noise types are used for training with a small amount of utterances and the noise types of the test set are the same as those of the training set. In this work, a large training set is formed by synthesizing the noisy speech data with a wide range of additive noises at different SNRs.

3.2 DNN-based Speech Enhancement

In [9], DNN was adopted as a regression model to predict the clean LPS features given the input noisy LPS features with acoustic context. This work improves the framework to predict the clean LPS and noise LPS features simultaneously in the output layer [10]. We believe the estimation of noise LPS will act as a regularization to the clean part. As for the DNN training, we first perform pre-training of a deep generative model with the LPS features of noisy speech by a stacking of multiple RBMs. Then the back-propagation with the MMSE-based objective function between the LPS features of the estimated and the reference (clean speech and noise) is adopted to train the DNN. Another two techniques, namely dropout training and noise-aware training (NAT) can be found in [19]. A stochastic gradient descent algorithm is performed in minibatches with multiple epochs to improve learning convergence as follows,

$$Er = \frac{1}{N}\sum_{n=1}^{N}(\beta\|\hat{\boldsymbol{X}}_n^{\text{clean}} - \boldsymbol{X}_n^{\text{clean}}\|_2^2 + (1-\beta)\|\hat{\boldsymbol{X}}_n^{\text{noise}} - \boldsymbol{X}_n^{\text{noise}}\|_2^2) \qquad (1)$$

where $\hat{\boldsymbol{X}}_n^{\text{clean}}$ and $\boldsymbol{X}_n^{\text{clean}}$ are the n^{th} D-dimensional vectors of estimated and reference clean features, respectively. In the same way, $\hat{\boldsymbol{X}}_n^{\text{noise}}$ and $\boldsymbol{X}_n^{\text{noise}}$ are the vectors of estimated and reference noise features. β is used to tune the

contribution from the speech part and the noise part. As the noise variance is large and not stable, we mainly focus on the speech part. The second term of Eq.(1) can be considered as a regularization term, which leads to a better generalization capacity for estimating the clean speech. Another benefit from the dual outputs DNN is the estimation of noise can be used in the following ideal ratio mask (IRM) based post-processing module:

$$\widehat{\boldsymbol{IRM}}_n(d) = \sqrt{\frac{\exp(\hat{\boldsymbol{X}}_n^{\text{clean}}(d))}{\exp(\hat{\boldsymbol{X}}_n^{\text{clean}}(d)) + \exp(\hat{\boldsymbol{X}}_n^{\text{noise}}(d))}} \tag{2}$$

Different from [6] where the IRM is directly predicted by a well trained IRM-DNN, the IRM here is estimated by the DNN output for each dimension d, which is used for post-processing as follows

$$\hat{\boldsymbol{X}}_n(d) = \begin{cases} \boldsymbol{Y}_n(d) & \widehat{\boldsymbol{IRM}}_n(d) > \gamma \\ \hat{\boldsymbol{X}}_n^{\text{clean}}(d) & \widehat{\boldsymbol{IRM}}_n(d) < \lambda \\ (\hat{\boldsymbol{X}}_n^{\text{clean}}(d) + \boldsymbol{Y}_n(d))/2 & \text{otherwise} \end{cases} \tag{3}$$

where, $\hat{\boldsymbol{X}}_n$ and $\boldsymbol{Y}_n$ are the vectors of final enhanced speech and noisy speech, respectively. γ and λ are the thresholds to improve the overall performance.

4 Experimental Results and Analysis

4.1 Experimental Setup

In [9], 104 noise types were used as the noise signals for synthesizing the noisy speech training samples. In this study, we add another home-made 200 hours real-world noises [4] to handle a wide range of additive noise in the real-world situations. 100 hours clean Mandarin data collected by iFlytek were added with the above-mentioned background noises and 5 levels of SNR, at 20dB, 15dB, 10dB, 5dB and 0dB, to build a multi-condition stereo training set. The whole 100-hour training data was used for baseline system and VAD model training. As for VAD training, the frame-level reference labels of each noisy utterance were generated by conventional VAD tool on the corresponding clean utterance. Then, we use the speech segments of the multi-condition training set (about 60 hours) for CSE model training. The training method is same with the baseline enhancement subsystem. The final joint DNN based SE system designed for low SNR environments was obtained under the framework illustrated in Fig. 2, denoted as JDNN-SE. Another 200 clean utterances covering 20 males and 17 females were used to construct the test set for each combination of noise types (NOISEX-92 corpus: babble and factory, real-recorded: mess hall and Karaoke

[4] The noise types are vehicle: bus, train, plane and car; exhibition hall; meeting room; office; emporium; family living room; factory; bus station; mess hall; KTV; musical instruments.

Table 1. *PESQ and STOI comparisons of four DNN-based SE systems averaged on the test sets for the four unseen noise conditions at different SNRs.*

Noise Type	SNR	PESQ				STOI			
		Noisy	Baseline	JDNN-SE	Oracle	Noisy	Baseline	JDNN-SE	Oracle
Babble	5dB	1.709	2.043	2.248	2.279	0.778	0.795	0.840	0.856
	0dB	1.341	1.307	1.732	1.802	0.678	0.603	0.717	0.758
	-5dB	1.057	0.793	1.126	1.174	0.567	0.396	0.557	0.606
Factory	5dB	1.594	1.990	2.300	2.353	0.778	0.761	0.839	0.861
	0dB	1.233	1.500	1.905	1.951	0.679	0.606	0.745	0.772
	-5dB	0.950	1.030	1.332	1.332	0.571	0.463	0.601	0.627
Mess Hall	5dB	1.655	2.048	2.286	2.280	0.787	0.809	0.854	0.863
	0dB	1.311	1.506	1.895	1.894	0.689	0.664	0.761	0.778
	-5dB	1.057	0.927	1.272	1.291	0.579	0.478	0.609	0.633
KTV	5dB	1.885	2.347	2.416	2.403	0.829	0.874	0.885	0.891
	0dB	1.526	1.939	2.077	2.066	0.754	0.796	0.824	0.835
	-5dB	1.198	1.394	1.595	1.619	0.665	0.672	0.726	0.741

Television (KTV)) and SNR levels (-5dB, 0dB, 5dB). All the noises, speakers and texts in test set are different from those in the training set.

For both the regression DNN and classification DNN, sigmoid activation function was used and the number of units in each hidden layer was set to 2048 by default. The mini-batch size N was set to 128. The regularization weighting coefficient β in Eq.(1) was 0.8. γ and λ in Eq.(3) were set to 0.75 and 0.1, respectively. The other tuning parameters of DNN were set according to [19, 20]. The half-window size τ for VAD smoothing was 5. The performance was evaluated using two measures, namely short-time objective intelligibility (STOI) [21] and perceptual evaluation of speech quality (PESQ) [22] measures.

4.2 Results and Analysis

Table 1 gives a performance comparison of different DNN-based SE systems for the four unseen noise environments with different SNRs averaged on the test set. Noisy means the original noisy speech without any processing. The difference between Oracle and JDNN-SE is whether they use clean reference VAD annotations in the enhancement stage. Compared with the noisy speech results, baseline system showed that the speech quality is very poor at SNR = -5dB, and the performance was not satisfactory at SNR = 0dB. Our proposed JDNN-SE system overwhelmed baseline at all SNRs, especially at low SNRs, e.g., 0.333 PESQ improvement and 0.161 STOI improvement at SNR = -5dB in babble noise environment. Finally, the gap between JDNN-SE and Oracle was small compared with that between Baseline and JDNN-SE. This implied that our DNN-based VAD was effective and robust to noise types. Fig. 3 presented spectrograms of an utterance. The improved DNN could enhance the speech

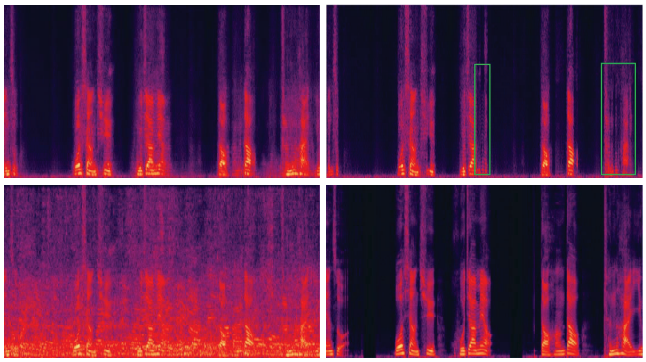

Fig. 3. *Four spectrograms of an utterance corrupted by babble noise at 0dB SNR: JDNN-SE system (upper left, PESQ = 2.115), DNN baseline (upper right, PESQ = 1.585), noisy (bottom left, PESQ = 1.602) and clean speech (bottom right, PESQ = 4.5).*

with less speech distortion, especially at the noisy speech segments which are similar to noise. More results can be found at the demo website [5].

5 Conclusion

We have proposed an improved speech enhancement framework to increase speech intelligibility in low SNR environments. In this method, speech and non-speech frames are presented to specific subsystem separately. With frame-level VAD prediction and corresponding soft decision fusion, we obtain the final enhanced speech. The proposed joint DNN based SE system can yield a significant improvement when compared with our baseline, especially in low SNR conditions. As for future work, we will focus on designing multiple DNNs with even more detailed resolution at various frame-level SNRs.

References

[1] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 27, no. 2, pp. 113–120, 1979.

[2] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 33, no. 2, pp. 443–445, 1985.

[3] I. Cohen, "Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging," *Speech and Audio Processing, IEEE Transactions on*, vol. 11, no. 5, pp. 466–475, 2003.

[4] N. Mohammadiha, P. Smaragdis, and A. Leijon, "Supervised and unsupervised speech enhancement using nonnegative matrix factorization," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 21, no. 10, pp. 2140–2151, 2013.

[5] `http://home.ustc.edu.cn/~gtian09/demos/LowSNR-SEDNN.html`

[5] Y. X. Wang and D. L. Wang, "Towards scaling up classification-based speech separation," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 21, no. 7, pp. 1381–1390, 2013.

[6] A. Narayanan and D. L. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," in *ICASSP*, 2013, pp. 7092–7096.

[7] J. Du and Q. Huo, "A speech enhancement approach using piecewise linear approximation of an explicit model of environmental distortions," in *INTERSPEECH*, 2008, pp. 569–572.

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

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

[10] Y.-H. Tu, J. Du, Y. Xu, L.-R. Dai, and C.-H. Lee, "Speech separation based on improved deep neural networks with dual outputs of speech features for both target and interfering speakers," in *ISCSLP*, 2014, pp. 250–254.

[11] J. Du, Q. Wang, T. Gao, Y. Xu, L.-R. Dai, and C.-H. Lee, "Robust speech recognition with speech enhanced deep neural networks," in *INTERSPEECH*, 2014, pp. 616–620.

[12] 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 *ICASSP*, 2015, accepted.

[13] A. Varga and H. J. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech communication*, vol. 12, no. 3, pp. 247–251, 1993.

[14] J. Sohn and W. Sung, "A voice activity detector employing soft decision based noise spectrum adaptation," in *ICASSP*, 1998, pp. 365–368.

[15] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *Signal Processing Letters, IEEE*, vol. 6, no. 1, pp. 1–3, 1999.

[16] X.-L. Zhang and J. Wu, "Deep belief networks based voice activity detection," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 21, no. 4, pp. 697–710, 2013.

[17] X.-L. Zhang and D. L. Wang, "Boosted deep neural networks and multi-resolution cochleagram features for voice activity detection," in *INTERSPEECH*, 2014, pp. 1534–1538.

[18] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

[19] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "Dynamic noise aware training for speech enhancement based on deep neural networks," in *INTERSPEECH*, 2014, pp. 2670–2674.

[20] G. Hinton, "A practical guide to training restricted Boltzmann machines," in *Neural Networks: Tricks of the Trade*, pp. 599–619. Springer, 2012.

[21] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *ICASSP*, 2010, pp. 4214–4217.

[22] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *ICASSP*, 2001, pp. 749–752.