Boosting DNN-Based Speech Enhancement via Explicit Transformations

Qing Wang, Jun Du and Li-Rong Dai
University of Science and Technology of China, Hefei, Anhui, China
E-mail: xiaosong@mail.ustc.edu.cn, {jundu,lrdai}@ustc.edu.cn

Abstract—**In this study, we investigate on the learning behaviors of DNN by explicit feature transformations. As a demonstration, linear and logarithm transformations, corresponding to the amplitude spectra and log-power spectra, are compared with the same minimum mean squared error (MMSE) objective function for optimizing DNN parameters. Based on the experimental analysis of the DNN learning behaviors, we make an interesting observation that the learning with the amplitude spectra tends to improve the speech intelligibility while the learning with the log-power spectra yields better speech quality. By leveraging on this strong complementarity, the feature concatenation with two transformations for the input layer and post-processing with two learned targets are proposed to boost DNN-based speech enhancement.**

I. Introduction

With the fast development of mobile internet, speech enhancement techniques have became extremely important in real-world applications, such as automatic speech recognition (ASR), mobile communication and hearing aids [1]. Historically, many signal processing methods for speech enhancement have been proposed during the past several decades, such as spectral subtraction [2], Wiener filtering [3], minimum mean squared error (MMSE) estimation [4], [5] and optimally-modified log-spectral amplitude (OM-LSA) speech estimator [6]. Model assumptions for the interactions between speech and noise were made in these methods, which could often lead to an imperfect listening quality of the enhanced speech. For example, most of these techniques can not make a good estimate of clean speech in highly non-stationary noise cases and musical noise artifacts [7] might be caused.

Recently, a bunch of deep learning based approaches were proposed for speech enhancement with promising results. In [8], stacked denoising autoencoder (SDA) based speech enhancement methods were adopted to model the complicated relationship between noisy speech and clean speech. In our recent work [9], [10], a novel speech enhancement framework via deep neural network (DNN) as a regression model to predict the clean log-power spectra (LPS) features from noisy LPS features was designed. No musical noise was found in the enhanced speech and highly non-stationary noise could be suppressed. Furthermore, large scale of noise types could be included in the training set to improve the generalization capacity to unseen noise environments [11].

However, according to our analysis, DNN-based speech enhancement in the log-power spectral domain tends to eliminate the noises with the risk of introducing speech distortions. Especially in low signal-to-noise ratio (SNR) conditions, it often leads to severe distortions in the speech segments. To address this problem, we revisit the feature design of the input noisy speech and output clean speech in the regression DNN learning. In a general framework, the final features fed to DNN can be generated via the amplitude spectra (AS) with an explicit transformation. The learning behavior of DNN varies with different transformations. Specifically, the linear and logarithm transformations, corresponding to AS features and LPS features are compared, which is similar to a prior work [4], [5] proposed by Ephraim and Malah in terms of the comparison between short-time spectral amplitude estimator and short-time log-spectral amplitude estimator. We make an interesting observation that the DNN learning in AS feature domain tends to improve the speech intelligibility while the DNN learning in the LPS feature domain yields better speech quality. By leveraging on this strong complementarity, feature concatenation with two transformations for the input layer and post-processing with two learned targets are proposed to improve the performance of the objective measures for both speech quality and speech intelligibility, namely perceptual evaluation of speech quality (PESQ) [12] and short-time objective intelligibility (STOI) [13].

II. System Overview

A blockdiagram of our proposed speech enhancement system is illustrated in Fig. 1. In the training stage, a regression DNN model is trained from a collection of stereo data, consisting of pairs of noisy and clean speech represented by transformed features from the amplitude spectra via the function $g(\cdot)$. To optimize the parameters of DNN, the MMSE criterion is adopted as follows,

$$E = \frac{1}{N}\sum_{n=1}^{N}\|\hat{\boldsymbol{X}}_n^g(\boldsymbol{Y}_n^g, \boldsymbol{W}, \boldsymbol{b}) - \boldsymbol{X}_n^g\|_2^2 + \kappa\|\boldsymbol{W}\|_2^2 \qquad (1)$$

where $\hat{\boldsymbol{X}}_n^g$ and $\boldsymbol{X}_n^g$ are the n^{th} D-dimensional vectors of estimated and clean reference features with the transformation function g, respectively. $\boldsymbol{Y}_n^g$ is the input noisy feature vector. $\boldsymbol{W}$ and $\boldsymbol{b}$ denote all the weight and bias parameters. κ is the regularization weighting coefficient to avoid over-fitting.

In the enhancement stage, the well-trained DNN model is fed with the noisy features to generate the enhanced features. The additional phase information is calculated from the original noisy speech. Finally an overlap-add method is used to

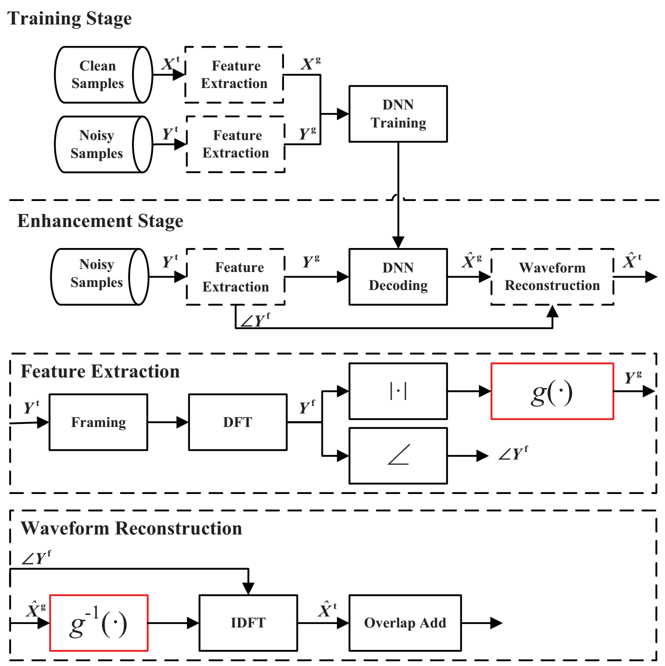

Fig. 1. A block diagram of the DNN-based speech enhancement via the explicit transformation.

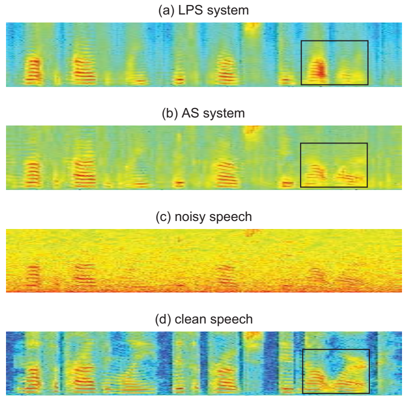

Fig. 2. Spectrogram of an utterance tested with pink noise at -5dB SNR: (a) DNN approach using LPS features (PESQ=1.810, STOI=0.645), (b) DNN approach using AS features (PESQ=1.680, STOI=0.733), (c) noisy speech (PESQ=1.107, STOI=0.573), (d) clean speech.

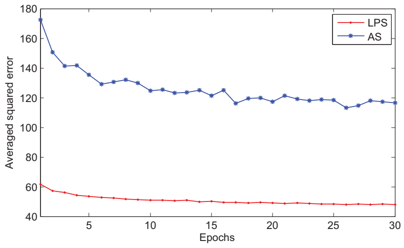

Fig. 3. The comparison of DNN learning curves using averaged squared errors on development set in LPS domain.

synthesize the waveform of the enhanced speech. A detailed description of DNN training, feature extraction, and waveform reconstruction can refer to [9], [14].

In our previous work [9], [10], only the DNN learning in the LPS feature domain is conducted, which corresponds to the logarithm transformation function for $g(\cdot)$. But in low SNR conditions, this regression DNN often focuses on the noise removal with the risk of introducing speech distortions and even speech information lost. This characteristic should be partly due to the feature design. To address this problem, the linear transformation function (actually an identity function) for $g(\cdot)$, corresponding to AS feature domain, is investigated. In the next section, we will make a comparison of the DNN learning behaviors between these two transformations.

III. The DNN Learning Behavior

First, we show an utterance example corrupted by pink noise at -5dB SNR in Fig. 2. The DNN approach using LPS features successfully eliminates most of the background noises with a better PESQ for speech quality. However the speech information is partially lost in the rectangle box, yielding a worse STOI for speech intelligibility. The observation for DNN approach using AS features is opposite, namely better speech preservation (a better STOI) with more residual noises (a worse PESQ). By this comparison, it seems that LPS and AS systems are strongly complementary to each other in terms of different evaluation measures.

The reason for this interesting observation can be explained by the DNN learning behaviors illustrated in Fig. 3 and Fig. 4. From the learning curves in Fig. 3, the LPS system always generates smaller squared errors than AS system in the LPS domain, which is reasonable as the objective function of DNN for LPS system exactly aims to minimize the mean square error in the LPS domain while the AS system is optimized in AS domain rather than LPS domain. And this learning curve can give a rough explanation to Fig. 2 as the smaller squared error can lead to less background noises in the enhanced speech. To have a deeper understanding of the DNN learning behavior, a higher resolution analysis as shown in Fig. 4 is given by using the distribution with respect to the frame-level squared error and frame-level SNR on the development set in LPS domain after DNN learning. Overall, the distribution of LPS system mainly focuses on the lower squared error area (marked red) than AS system, which is similar to learning curves in Fig. 3. Specifically, in the high frame-level SNR range from 15dB to 20dB, larger squared errors are distributed in AS system than LPS system, which implies that AS system can not well handle the high SNR segments with more residual noises after enhancement. However, in the low frame-level SNR range below 5dB, we make an opposite observation that more squared errors are generated in LPS system, which indicates that LPS system can not well handle the low SNR segments with more speech distortions after enhancement.

IV. Improved DNN via Multiple Transforms

Based on the analysis in Section III, we aim at improving our previous DNN approach operating in LPS domain [9] by leveraging the strong complementarity between the two

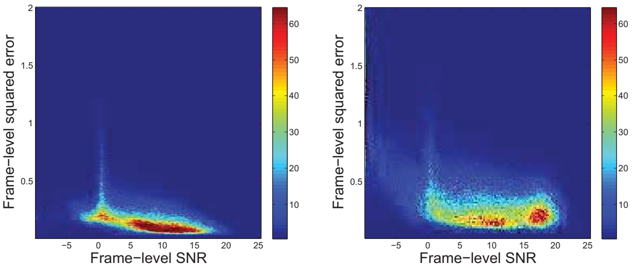

Fig. 4. The comparison of distribution with respect to frame-level squared error and frame-level SNR on the development set in LPS domain after DNN learning for LPS system (left) and AS system (right).

transformations corresponding to LPS and AS features in this section. Two strategies, namely feature concatenation and post-processing, are elaborated as follows.

A. Feature Concatenation

The feature design of the regression DNN for denoising plays an important role for both input layer and output layer. Feature concatenation in the input layer is one simple way to fully utilize multiple feature transformations. Each transformation might characterize one critical property of the speech signal. With different learning targets, two concatenation systems can be designed as in Tab. I. The LPS+AS-LPS is the concatenating version of LPS-LPS with AS features appended while LPS+AS-AS corresponds to AS-AS augmented with LPS features. The effectiveness of feature combination has been demonstrated in many speech areas. For example in speech recognition, feature combination may lead to significant performance improvements [15]. One similar work in [16] shows the effectivity of feature concatenation in DNN-based speech separation. It is expected that the concatenation of LPS and AS features can improve both measures for speech quality and speech intelligibility.

B. Post-processing

For the two concatenation systems, namely LPS+AS-LPS and LPS+AS-AS, although the input features are the same, the final enhancement results should vary a lot due to the different learning targets. So a post-processing approach to leverage on the outputs of both LPS+AS-LPS and LPS+AS-AS DNNs is proposed as follows,

$$\hat{\boldsymbol{X}}^{\mathrm{PP}} = \alpha * \hat{\boldsymbol{X}}^{\mathrm{LPS}} + (1-\alpha)\hat{\boldsymbol{X}}^{\mathrm{AS}\rightarrow\mathrm{LPS}} \tag{2}$$

where $\hat{\boldsymbol{X}}^{\mathrm{LPS}}$ is the output of LPS+AS-LPS DNN system while $\hat{\boldsymbol{X}}^{\mathrm{AS}\rightarrow\mathrm{LPS}}$ is the transformed version from the output of LPS+AS-AS DNN system to the LPS domain. α is the weighting coefficient.

V. Experiments and Results Analysis

In this study, 115 noise types including 100 noise types recorded by G. Hu [17] and some other musical noises were adopted to improve the generalization capacity of DNN. The clean speech data is derived from the TIMIT corpus [18]. All 4620 utterances from the training set of the TIMIT database were corrupted with the abovementioned 115 noise types at six levels of SNR, i.e., 20dB, 15dB, 10dB, 5dB, 0dB, and -5dB, to build 80-hour multi-condition training set, consisting of pairs of clean and noisy speech utterances. The 192 utterances from core test set of TIMIT database were used to construct the test set for each combination of noise types and SNR levels. As we only conducted the evaluation of mismatched noise types, 13 unseen noise types[1], from the NOISEX-92 corpus [19], were adopted for testing.

TABLE I
Four DNN systems.

DNN System	Input Feature	Output Target
LPS-LPS	LPS	LPS
AS-AS	AS	AS
LPS+AS-LPS	LPS+AS	LPS
LPS+AS-AS	LPS+AS	AS

As for signal analysis, all experiments were conducted on waveforms with 16kHz sample rate, and the corresponding frame length was set to 512 samples (or 32 msec) with a frame shift of 256 samples. A short-time Fourier analysis was used to compute the DFT of each overlapping windowed frame. Then 257-dimensional LPS features [14] or AS features were used to train DNNs. PESQ and STOI were used to assess the quality and intelligibility of the enhanced speech.

All DNN configurations were fixed at $L = 3$ hidden layers, 2048 units at each hidden layer, and 7-frame acoustic context. Rectified linear units (ReLU) [20] was used as the activation function of DNN, and the DNN was initialized with random weights. Dropout and static noise aware training [21] were used to improve its generalization capacity for unseen noise types. Other details of the setup can be found in [10].

A. Experiments on Feature Concatenation

Tab. II presents the average STOI and PESQ comparison for four DNN systems in Tab. I on the test set at different SNRs of the 13 unseen noise environments. Better LSD and PESQ performances could be obtained by LPS-LPS system, while better STOI and SSNR performances were achieved by AS-AS system. This has been partially interpreted by the learning curves of Fig. 3 and the distributions of Fig. 4 for PESQ and STOI in Section III. In other words, LPS-LPS system could bring better speech quality in terms of PESQ especially for high SNR cases while AS-AS system tended to improve speech intelligibility in terms of STOI for low SNR cases. For example, PESQ of LPS-LPS system was 3.58 at 20dB while PESQ of AS-AS system was 3.24. On the contrary, STOI of LPS-LPS system was 0.697 at -5dB while STOI of AS-AS system was 0.739.

By conducting feature concatenation for input layer, all four evaluation metrics were improved for both feature transformations, corresponding to LPS+AS-LPS system and LPS+AS-AS

[1] The 13 unseen environment noises for evaluation are Buccaneer1, Buccaneer2, Destroyer engine, Destroyer ops, F16, Factory1, Factory2, HF channel, Leopard, M109, Machine gun, Pink, and Volvo. They are all collected from the NOISEX-92 corpus.

TABLE II

Average STOI and PESQ comparison at different SNRs on 13 unseen noise environments, among: LPS-LPS system, LPS+AS-LPS system, AS-AS system, and LPS+AS-AS system.

	LPS-LPS		LPS+AS-LPS		AS-AS		LPS+AS-AS	
SNR(dB)	STOI	PESQ	STOI	PESQ	STOI	PESQ	STOI	PESQ
20	0.960	3.58	0.963	3.62	0.969	3.24	0.972	3.46
15	0.946	3.37	0.949	3.41	0.955	3.10	0.958	3.28
10	0.920	3.12	0.924	3.16	0.930	2.92	0.933	3.06
5	0.875	2.83	0.884	2.88	0.890	2.70	0.893	2.81
0	0.801	2.49	0.822	2.56	0.828	2.42	0.833	2.52
-5	0.697	2.09	0.730	2.17	0.739	2.08	0.746	2.19
Avg	0.866	2.91	0.879	2.97	0.885	2.74	0.889	2.89

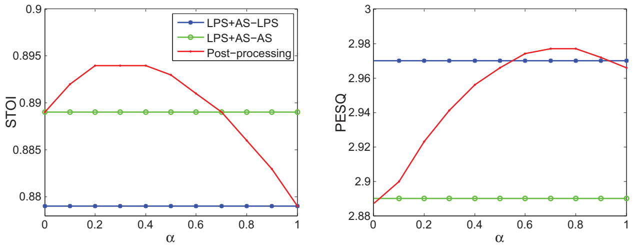

Fig. 5. Average STOI and PESQ performance on 13 unseen noise types across different SNRs.

system. And the gap between the two concatenated systems was smaller than that between two baseline systems. The STOI performance of LPS+AS-LPS system was improved over LPS-LPS system, from 0.866 to 0.879 in average. And significant improvement was achieved at lower SNRs, e.g., from 0.697 to 0.730 at -5dB SNR. And PESQ performance of LPS+AS-AS system was also improved over AS-AS system, from 2.74 to 2.89 in average. All those results demonstrated the strong complementarity between the two feature transformations.

B. *Experiments on Post-processing*

On top of LPS+AS-LPS and LPS+AS-AS systems, the post-processing was conducted via (2). Fig. 5 lists average STOI and PESQ performance on the test set of 13 unseen noise types across different SNRs. With the weighting factor α ranging from 0 to 1, the PESQ and STOI performance was not monotonically increasing nor decreasing between the two systems, but with peak values. This confirmed that the two concatenation systems could still be complementary. And both the optimal values of α for STOI and PESQ were close to 0.5.

VI. Conclusions

In this paper, we first analyze the learning behavior of DNN-based speech enhancement with LPS features and AS features. Experimental results show that these two feature transformations can improve the quality and intelligibility of the enhanced speech, respectively. Thus, we propose two approaches to boost DNN-based speech enhancement by leveraging on the complementarity between these two features. With feature concatenation for the input layer, evaluation metrics were improved. Intelligibility of enhanced speech was significantly improved especially at low SNRs. By post-processing with two learned targets, the performance can be further improved.

Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grants No. 61305002.

References

[1] J. Benesty, S. Makino, and J. D. Chen, *Speech Enhancement*. Springer, 2005.

[2] 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.

[3] J. S. Lim and A. V. Oppenheim, “All-pole modeling of degraded speech,” *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 26, no. 3, pp. 197–210, 1978.

[4] Y. Ephraim and D. Malah, “Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator,” *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 32, no. 6, pp. 1109–1121, 1984.

[5] ——, “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.

[6] I. Cohen and B. Berdugo, “Speech enhancement for non-stationary noise environments,” *Signal Processing*, vol. 81, no. 11, pp. 2403–2418, 2001.

[7] A. Hussain, M. Chetouani, S. Squartini, B. A., and P. F., *Nonlinear speech enhancement: An overview*. Springer, 2007.

[8] X. Lu, Y. TSao, S. Matsuda, and C. Hori, “Speech enhancement based on deep denoising autoencoder,” in *INTERSPEECH*, 2013, pp. 436–440.

[9] 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, Jan 2014.

[10] ——, “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, Jan 2015.

[11] Y. Xu, J. Du, Z. Huang, L.-R. Dai, and C.-H. Lee, “Multi-objective learning and mask-based post-processing for deep neural network based speech enhancement,” in *INTERSPEECH*, 2015, pp. 1508–1512.

[12] 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 *Acoustics, Speech and Signal Processing (ICASSP), 2001 IEEE International Conference on*, vol. 2. IEEE, 2001, pp. 749–752.

[13] C. Taal, R. Hendriks, R. Heusdens, and J. Jensen, “A short-time objective intelligibility measure for time-frequency weighted noisy speech,” in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, 2010, pp. 4214–4217.

[14] 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.

[15] G. Garau and S. Renals, “Combining spectral representations for large-vocabulary continuous speech recognition,” *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 16, no. 3, pp. 508–518, 2008.

[16] Y. X. Wang, K. Han, and W. D. L., “Exploring monaurl features for classification-based speech segregation,” *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 21, no. 2, pp. 270–279, 2013.

[17] G. Hu, “100 nonspeech environmental sounds, 2004.”

[18] J. S. Garofolo, “Getting started with the darpa timit cd-rom: An acoustic phonetic continuous speech database,” NIST, Tech. Rep., 1988.

[19] 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.

[20] G. Dahl, T. Sainsth, and G. Hinton, “Improving deep neural networks for lvcsr using rectified linear units and dropout,” in *Acoustics Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, 2013, pp. 8609–8613.

[21] M. Seltzer, D. Yu, and Y. Wang, “An investigation of deep neural networks for noise robust speech recognition,” in *Acoustics Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, 2013, pp. 7398–7402.