A UNIFIED SPEAKER-DEPENDENT SPEECH SEPARATION AND ENHANCEMENT SYSTEM BASED ON DEEP NEURAL NETWORKS

Tian Gao[1]*, Jun Du*[1]*, Li Xu*[2]*, 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@mail.ustc.edu.cn, {jundu,lrdai}@ustc.edu.cn
{lixu,congliu2}@iflytek.com, chl@ece.gatech.edu

ABSTRACT

Speech enhancement and speech separation are important frontends of many speech processing systems. In real tasks, the background noises are often mixed with some human voice interferences. In this paper, we explore a framework to unify speech enhancement and speech separation for a speaker-dependent scenario based on deep neural networks (DNNs). Using a supervised method, DNN is adopted to directly model a nonlinear mapping function between noisy and clean speech signals. The signals of speaker interferers are considered as one type of universal noise signals in our framework. In order to be able to handle a wide range of additive noise in the real-world situations, a large training set that encompasses many possible combinations of speech and noise types, is designed. Experimental results demonstrate that the proposed framework can get the comparable performances to those single speech enhancement or separation systems. Furthermore, the resulting DNN model, trained with artificial synthesized data, is also effective in dealing with noisy speech data recorded in real-world conditions.

Index Terms— speech enhancement, speech separation, speaker-dependent, deep neural networks, supervised method

1. INTRODUCTION

Speech enhancement and speech separation are important frontends of many speech processing systems such as speech communication and automatic speech recognition [1, 2]. The goal of speech enhancement is to improve the intelligibility and quality of a noisy speech signal degraded in adverse conditions. Similar with speech enhancement, speech separation aims at separating the voice of each speaker when multiple speakers talk simultaneously [3, 4]. Considering the process of noise corruption on speech is very complicated, the enhancement and separation performance is still unsatisfactory.

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

Numerous speech enhancement methods were developed over the past several decades, such as spectral subtraction [5], Wiener filtering [6], minimum mean squared error (MMSE) estimation [7, 8] and optimally-modified log-spectral amplitude (OM-LSA) speech estimator [9, 10]. In most of these algorithms, it is assumed that an estimate of the noise spectrum is available [11]. Its noise tracking capacity is limited for highly non-stationary noise cases, and it tends to distort the speech component in mixed signals if it is tuned for a better noise reduction.

Recently, supervised learning methods are becoming more popular. Some data-driven methods attempt to make a binary classification decision on time-frequency (T-F) units, such as estimating the ideal binary mask for monaural speech separation [12], however the acoustic context information of the T-F unit is not well utilized in a classification framework. In [13], DNNs were used to estimate a smoothed ideal ratio mask (IRM) in the Mel frequency domain for robust ASR. In [14], a regression DNN-based speech enhancement framework via training a deep and wide neural network architecture using a large collection of heterogeneous training data was proposed. It was found multi-condition training with many kinds of noise types can achieve a good generalization capability to unseen noise environments. Moreover, the proposed DNN framework is also powerful to cope with non-stationary noises in real-world environments.

From a unified viewpoint, both speech enhancement and separation aim at removing interference sources. Generally, it is hard to use one single system to handle both background noises and speaker interferers with conventional approaches. However, if the target speech to be separated is from a specific speaker, speech enhancement and separation could be unified. The unified system is meaningful because the real-world noises are often mixed with some human voice interferences. In this paper, based on our previous work [15, 16], we propose a unified speaker-dependent speech separation and enhancement framework. For convenience, we call this task as unified speech enhancement (U-SE). The main contributions of this

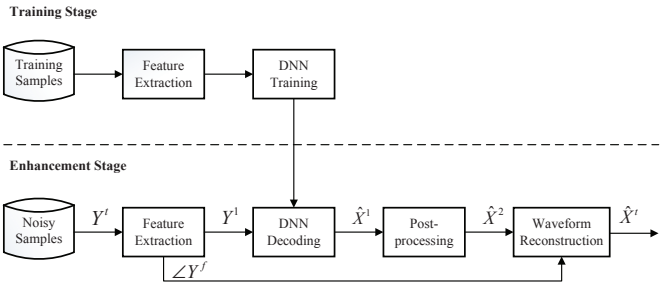

Fig. 1. Overall development flow and architecture.

paper are summarized as follows: (i) We propose a generalized framework of speech separation and enhancement for a target speaker. (ii) A large training set of Mandarin speech that encompasses many possible combinations of speech and noise types is designed. (iii) We employ an improved DNN architecture with dual outputs of speech features for both target and interference source in the output layer. And a novel post-processing with IRM is proposed. Empirical results demonstrate that the proposed framework can get the comparable performances to those single speech enhancement or separation systems for both synthetic and real-world conditions.

2. SYSTEM OVERVIEW

The overall flowchart of our proposed unified speech enhancement framework 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 the log-power spectra (LPS) features. In the enhancement stage, the well-trained DNN model is fed with the features of noisy speech in order to generate the enhanced LPS features. Another module, namely post-processing with IRM is proposed to improve the overall performance. The additional phase information is calculated from the original noisy speech. Finally an overlap-add method is used to synthesize the waveform of the estimated clean speech. A detailed description of the feature extraction module and waveform reconstruction module can be found in [17]. In the next section, the details of DNN-based speech enhancement are elaborated.

3. DNN-BASED SPEECH ENHANCEMENT

3.1. DNN Training

In [15], DNN was adopted as a regression model to predict the clean LPS features given the input noisy LPS features with acoustic context. The current work improves the framework to predict the clean LPS and noise LPS features simultaneously in the output layer as shown in Fig. 2. We believe the estimation of noise LPS will act as a regularization to the

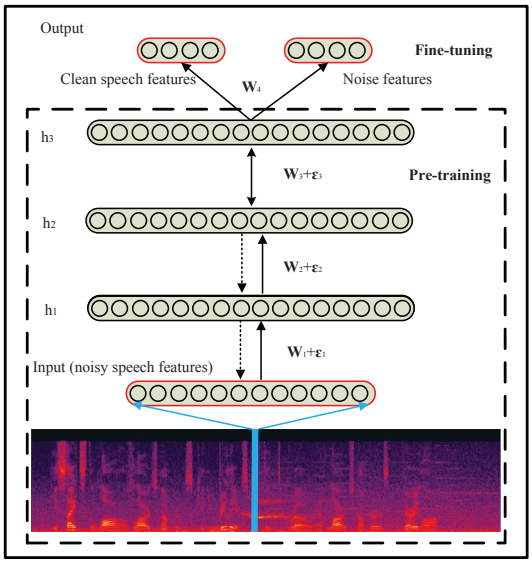

Fig. 2. DNN-based speech enhancement.

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 restricted Boltzmann machines (RBMs) [18]. Then the back-propagation with the MMSE-based object 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,

$$Er = \frac{1}{N}\sum_{n=1}^{N}(\alpha\|\hat{\boldsymbol{X}}_n^t - \boldsymbol{X}_n^t\|_2^2 + (1-\alpha)\|\hat{\boldsymbol{X}}_n^i - \boldsymbol{X}_n^i\|_2^2) \quad (1)$$

where $\hat{\boldsymbol{X}}_n^t$ and $\boldsymbol{X}_n^t$ are the n^{th} D-dimensional vectors of estimated and reference clean features of the target speaker, respectively. In the same way, $\hat{\boldsymbol{X}}_n^i$ and $\boldsymbol{X}_n^i$ are the vectors of estimated and reference noise features. α 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 target speech. Another benefit from the dual outputs DNN is the estimation of noise can be used in the following post-processing module.

3.2. Post-processing with IRM

Different from [13] where the IRM is directly predicted by a well trained IRM-DNN, the IRM here is estimated by the

DNN output for each dimension as follows,

$$\widehat{\boldsymbol{IRM}}_n(d) = \sqrt{\frac{\exp(\hat{\boldsymbol{X}}^t_n(d))}{\exp(\hat{\boldsymbol{X}}^t_n(d)) + \exp(\hat{\boldsymbol{X}}^i_n(d))}} \quad (2)$$

which is used in post-processing as follows,

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

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

4. EXPERIMENTAL RESULTS AND ANALYSIS

4.1. Experimental configurations

In [15], 104 noise types [1] were used as the noise signals for synthesizing the noisy speech training samples. In this study, we add another 200 hours real-world noises [2] to handle a wide range of additive noise in the real-world situations. 6 hours interference speech covering 30 males and 30 females are also used for speech separation. On the other hand, 2 hours Hi-Fi Mandarin data were recorded by the target male speaker as our clean data. The 2 hours clean data were added with the above-mentioned background noises and interference speech and 5 levels of Signal Noise Ratio (SNR), at 20dB, 15dB, 10dB, 5dB and 0dB, to build a multi-condition stereo training set. This resulted in a collection of about 100 hours of noisy training data (including two subsets, 80 hours for speech enhancement and the rest 20 hours for speech separation) used to train the DNN models. The whole 100-hour training data was used for a unified speech enhancement model training, namely U-SE. The two training subsets were used for training of individual speech enhancement and speech separation systems, namely SE and SS systems. Another 50 utterances recorded from the target speaker were used to construct the test set for each combination of noise types (KTV, mess hall, female and male interferers) and SNR levels (-5dB, 0dB, 5dB, 10dB and 15dB). It should be noted that all the noises in test set are different from those in the training set.

As for signal analysis, speech waveform was down-sampled to 16KHz, and the corresponding frame length was set to 512 samples (or 32 msec) with a frame shift was set to 256 samples. A short-time Fourier analysis was used to compute the DFT of each overlapping windowed frame. Then the 257-dimensional LPS features were used to train DNNs. The performance was evaluated using two measures, namely a short-time objective intelligibility (STOI) [20] and Perceptual evaluation of speech quality (PESQ) [21]. STOI is shown to be highly correlated to human speech intelligibility while PESQ has a high correlation with subjective scores and it ranges from -0.5 to 4.5.

The number of epoch for each layer of RBM pre-training was 20. Learning rate of pre-training was 0.0005. As for the finetuning, learning rate was set at 0.1 for the first 10 epochs, then decreased by 10% after every epoch. Total number of epoch was 30. The mini-batch size was set to N=128. Input features of DNNs were normalized to zero mean and unit variance. The DNN architecture was 2056-2048-2048-2048-514, which denoted that the size was 2056 (257*7+257, τ=3) at the input layer, 2048 for three hidden layers, and 514 for the output layer (dual outputs). The regularization weighting coefficient α in Eq.(1) was 0.8. β and λ in Eq.(3) were set to 0.75 and 0.1, respectively. Other details of the setup can be found in [19].

4.2. Results and Analysis

The performance (PESQ) of different systems on speech enhancement and separation task under different SNRs is shown in Table 1. Noisy, SE, SS and U-SE stand for original noisy speech, speech enhancement model (80-hour), speech separation model (20-hour), unified speech enhancement (100-hour), respectively. At first, with the comparison of SS and SE on speech enhancement task and the comparison of SE and SS on speech separation task as our cross validation, we observe that the performance of the cross testing is dramatically degraded, namley SE model on SS test data or SS model on SE test data. Then, the comparison of SE/SS and our proposed U-SE is the main focus of this work. The results show that the unified system can get fairly effect compared with the corresponding best single systems. In detail, SE is better than U-SE only at low SNRs on enhancement task. This is reasonable as the U-SE model aims to remove not only the background noises but also the interference speech. Different from enhancement task, it is interesting to find out that the U-SE is better than SS on separation task. This can be explained as the existence of 80-hour noisy data with background noises is helpful for speech separation.

Table 2 shows the STOI results among SE, SS, and U-SE on the speech enhancement and separation task under different SNRs. The conclusion is almost the same as PESQ results in Table 1. In summary, through the comparison of our proposed unified system (U-SE) and corresponding subsystems (SE, SS), the framework of speaker-dependent speech separation and enhancement is feasible

[1] The 104 noise types are N1-N17: Crowd noise; N18-N29: Machine noise; N30-N43: Alarm and siren; N44-N46: Traffic and car noise; N47-N55: Animal sound; N56-N69: Water sound; N70-N78: Wind; N79-N82: Bell; N83-N85: Cough; N86: Clap; N87: Snore; N88: Click; N88-N90: Laugh; N91-N92: Yawn; N93: Cry; N94: Shower; N95: Tooth brushing; N96-N97: Footsteps; N98: Door moving; N99-N100: Phone dialing; N101: AWGN, N102: Babble, N103: Restaurant, N104: Street.

[2] The noise types are Vehicle: bus, train, plane and car; Exhibition hall; Meeting room; Office; Emporium; Family living room; Factory; Bus station; Mess hall.

Table 1. Average PESQ comparison of different systems on speech enhancement and separation task under different SNRs on the test set (Noise types: KTV, mess hall, female and male interferers with target speaker).

	Speech Enhancement Task				Speech Separation Task			
System	Noisy	SS	SE	U-SE	Noisy	SE	SS	U-SE
SNR15	2.90	3.12	3.34	3.34	2.97	3.25	3.33	3.39
SNR10	2.63	2.87	3.12	3.12	2.69	2.98	3.10	3.15
SNR5	2.35	2.56	2.86	2.85	2.39	2.68	2.81	2.88
SNR0	2.05	2.21	2.52	2.47	2.04	2.31	2.51	2.56
SNR-5	1.76	1.86	2.10	1.94	1.67	1.91	2.21	2.19
Avg.	2.34	2.52	2.79	2.75	2.35	2.63	2.79	2.83

Table 2. Average STOI comparison of different systems on speech enhancement and separation task under different SNRs on the test set (Noise types: KTV, mess hall, female and male interferers with target speaker).

	Speech Enhancement Task				Speech Separation Task			
System	Noisy	SS	SE	U-SE	Noisy	SE	SS	U-SE
SNR15	0.82	0.86	0.87	0.87	0.83	0.86	0.88	0.88
SNR10	0.78	0.82	0.85	0.85	0.78	0.82	0.85	0.85
SNR5	0.71	0.77	0.82	0.82	0.70	0.77	0.80	0.81
SNR0	0.62	0.69	0.76	0.75	0.62	0.69	0.75	0.74
SNR-5	0.53	0.58	0.66	0.62	0.53	0.60	0.68	0.67
Avg.	0.69	0.75	0.79	0.78	0.69	0.75	0.79	0.79

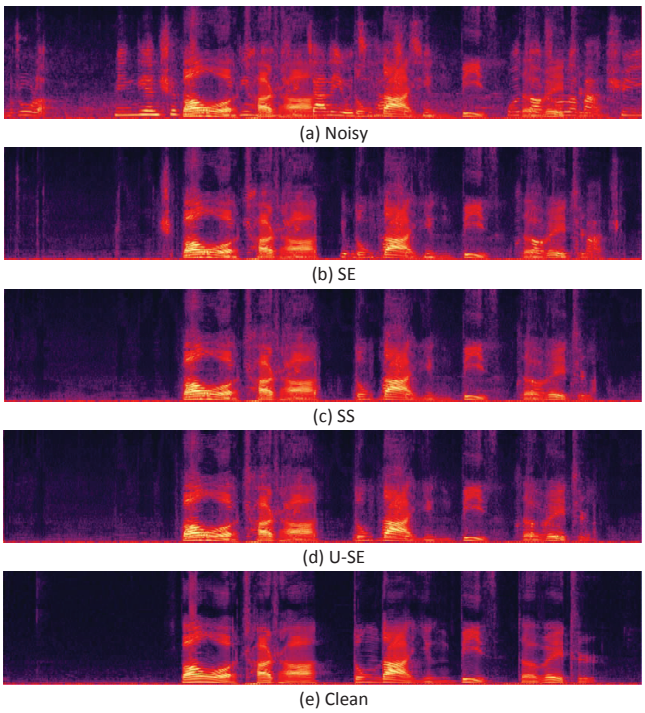

Fig. 3. Spectrograms of an utterance example corrupted by a female interferer at SNR=5dB: (a) Noisy speech, (b) SE enhanced, (c) SS enhanced, (d) U-SE enhanced and (e) clean speech.

and effective. The spectrograms of an processed example were presented in Fig. 3. More results could be found at http://home.ustc.edu.cn/˜gtian09/demos/USE_DNN.html.

5. CONCLUSION

In this paper, we employ a speaker-dependent framework to unify speech enhancement and speech separation based on deep neural networks (DNNs). We found that the DNN-based unified speech separation and enhancement system was effective to handle both speech enhancement and separation tasks. Two strategies, namely dual outputs and IRM-based post-processing were proposed and achieved a better performance. Moreover, a large training data of Mandarin speech with many noise types and combinations could achieve a good generalization capability to real-world noise environments. Empirical results demonstrate that the proposed framework can get fairly effect compared with the corresponding best single system in both synthetic and real-world conditions. In summary, the unified speech enhancement system for a speaker-dependent scenario is feasible and effective.

6. ACKNOWLEDGMENT

We thank iFLYTEK Research for providing the training data and DNN training platform.

7. REFERENCES

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

[2] D.L. Wang and Guy J. Brown, "Separation of speech from interfering sounds based on oscillatory correlation," *Neural Networks, IEEE Transactions on*, vol. 10, no. 3, pp. 684–697, May 1999.

[3] Guoning Hu and DeLiang Wang, "A tandem algorithm for pitch estimation and voiced speech segregation," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 18, no. 8, pp. 2067–2079, 2010.

[4] A.M. Reddy and B. Raj, "Soft mask methods for single-channel speaker separation," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 6, pp. 1766–1776, 2007.

[5] 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, Apr 1979.

[6] J.S. Lim and A.V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proceedings of the IEEE*, vol. 67, no. 12, pp. 1586–1604, Dec 1979.

[7] 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, Dec 1984.

[8] 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, Apr 1985.

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

[10] 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, Sept 2003.

[11] S. Rangachari and P. C. Loizou, "A noise-estimation algorithm for highly non-stationary environments," *Speech Communication*, vol. 48, no. 2, pp. 220–231, 2006.

[12] Yuxuan Wang and DeLiang Wang, "Towards scaling up classification-based speech separation," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 21, no. 7, pp. 1381–1390, July 2013.

[13] A. Narayanan and DeLiang 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*, May 2013, pp. 7092–7096.

[14] Yong Xu, Jun Du, Li-Rong Dai, and Chin-Hui 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.

[15] 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, Jan 2015.

[16] Yanhui Tu, Jun Du, Yong Xu, Lirong Dai, and Chin-Hui Lee, "Speech separation based on improved deep neural networks with dual outputs of speech features for both target and interfering speakers," in *Proceedings of the 9th International Symposium on Chinese Spoken Language Processing, ISCSLP 2014*, 2014, pp. 250–254.

[17] Jun Du and Qiang Huo, "A speech enhancement approach using piecewise linear approximation of an explicit model of environmental distortions," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2008, pp. 569–572.

[18] Yoshua Bengio, "Learning deep architectures for AI," *Foundations and trends in Machine Learning*, vol. 2, no. 1, pp. 1–127, 2009.

[19] Yong Xu, Jun Du, Li-Rong Dai, and Chin-Hui Lee, "Dynamic noise aware training for speech enhancement based on deep neural networks," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2014, pp. 2670–2674.

[20] C.H. Taal, R.C. 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.

[21] Recommendation P.862, "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," *International Telecommuni-cation Union-Telecommunication Standardisation Sector*, 2001.