# CROSS-DOMAIN COOPERATIVE DEEP STACKING NETWORK FOR SPEECH SEPARATION

*Wei Jiang*[1], *Shan Liang*[1], *Like Dong*[2], *Hong Yang*[2], *Wenju Liu*[1], *and Yunji Wang*[3]

[1]NLPR, Institute of Automation, Chinese Academy of Sciences, Beijing, China
[2]Electric Power Research Institute of ShanXi Electric Power Company, China State Grid Corp
[3]Electrical and Computer Engineering Department, The University of Texas at San Antonio, USA
{wjiang, sliang, lwj}@nlpr.ia.ac.cn, {likedong.csgc, hongyang.csgc, yunjiwang}@gmail.com

## ABSTRACT

Nowadays supervised speech separation has drawn much attention and shown great promise in the meantime. While there has been a lot of success, existing algorithms perform the task only in one preselected representative domain. In this study, we propose to perform the task in two different time-frequency domains simultaneously and cooperatively, which can model the implicit correlations between different representations of the same speech separation task. Besides, many time-frequency (T-F) units are dominated by noise in low signal-to-noise ratio (SNR) conditions, so more robust features are obtained by stacking features of original mixtures with that extracted from separated speech of each deep stacking network (DSN) block, which can be regarded as a denoised version of the original features. Quantitative experiments show that the proposed cross-domain cooperative deep stacking network (DSN-CDC) has enhanced modeling capability as well as generalization ability, which outperforms a previous algorithm based on standard deep neural networks.

***Index Terms***— Speech separation, cross-domain cooperative structure, deep stacking network, deep neural network

## 1. INTRODUCTION

Speech separation has been drawing more and more attention in speech processing society for years. Despite much progress has been made, separating speech from background noise at very low SNR is still very challenging, and traditional speech enhancement methods encounter much difficulty in this situation especially in single-channel systems. Meanwhile, the technique has wide applications in real life, such as robust automatic speech recognition (ASR) and hearing aid design. In this study, we focus on monaural speech separation from non-speech background noise at low SNR.

Nowadays, speech separation has been formulated as a supervised learning problem more frequently. With the popularity of masking methods in speech separation and denoising systems, various discriminative models are trained to predict a variety of ideal T-F masks that are treated as training targets [1, 2, 3]. Much efforts have been made in this field recently. For example, in order to extract the most effective features from the original mixtures, the performances of a variety of features were investigated [4, 5]. There are time and frequency correlations of T-F bins in any T-F representation of speech signals, so mask values of all the frequency bands of a frame are usually predicted together instead of predicting the mask value of one T-F unit independently [6]. To model the time coherence, multiple frames expansion is generally utilized, and a model named deep stacking network with time series (DSN-TS) has also been proposed [7]. Besides, due to the excellent power of feature extraction and transformation, deep neural networks, more often than not, has been adopted as the corresponding classifier or regressor of speech separation systems based on supervised learning.

However, there are two defects in previous speech separation systems. First, they typically perform speech separation in one preselected domain. Spectrogram and cochleagram are two commonly used time-frequency domains nowadays. Compared with each other, a cochleagram provides a much higher frequency resolution at low frequencies, while a spectrogram has a better resolution at high frequencies [8]. When separation is carried out in one domain, complementary information that could be obtained from the other representation is neglected. Second, acoustic features and transformed features (eg. learned with autoencoders) are extracted from the original mixture signals and a one-round separation process is performed in previous algorithms. However, we could expect to extract more reliable features from the separated speech, noting the separated speech is more close to clean speech. In other words, there is information in the separated speech that can be utilized to further improve the separation.

To address the deficiencies proposed above, we propose a cross-domain cooperative deep stacking network (DSN-CDC) structure to perform speech separation. This model could capture the complementary information between different time-frequency representations and a denoised version

This work was supported in part by the China National Nature Science Foundation (No.91120303, No.90820011, No.90820303 and No.61403370).

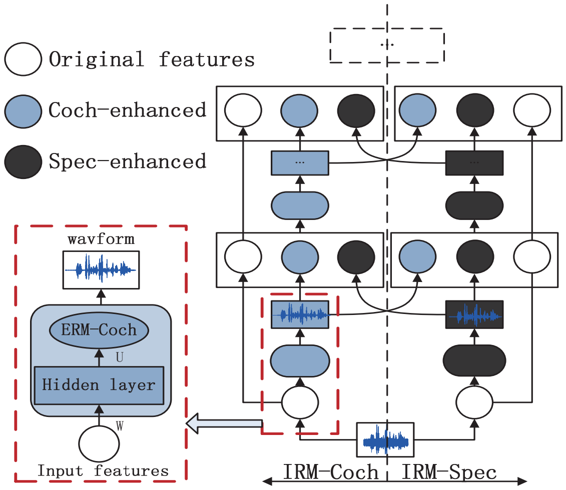

**Fig. 1**. Architecture of the network. Coch-enhanced /spec-enhanced indicates features extracted from separated speech in the cochleagram /spectrogram domain. IRM-Coch /IRM-Spec means that we take ideal ratio mask (IRM) in the cochleagram /spectrogram domain as the training target of the left /right DSN. ERM means estimated ratio mask.

of features could be directly obtained from the separated speech of the previous DSN block. Due to these two reasons, better performance is achieved by the model. Besides, as there is no need to use denoising autoencoders (DA) [9, 10] or denoising MLP [2] in the proposed structure, where only one hidden layer in each block of DSN is used, our model is much easier to train than previous speech separation systems based on deep neural networks (DNN).

The paper is organized as follows. In Section 2, the model structure of the proposed method is described. Experiments and analysis are presented in Section 3. Section 4 concludes the paper.

## 2. METHOD DESCRIPTION

### 2.1. An Architectural Overview

The architecture of the network is shown in Fig. 1. As can be seen from the figure, two DSNs are used to do speech separation in two time-frequency domains (cochleagram and spectrogram) simultaneously and cooperatively. Cooperation here means that "input" expanded vectors not only include features from the previous layer, but also features extracted from separated speech in the other domain. We expect preliminarily separated speech from the two domains could provide complementary information for the separation process of the next layer. In our experiments, obvious improvements are obtained in both domains. Thus we name the network structure proposed here as cross-domain cooperative deep stacking network (DSN-CDC).

It is worth noting that a variant of feature expansion practice is used in this study, where the output of each block is not directly used for stacking. Instead, the output of each block, which is the estimated ratio mask (ERM), is first used to obtain the time-domain separated speech, and then features extracted from the separated speech is concatenated with the original features. In some sense, this practice makes our model look like a tandem or cascading system, as the separated speech could be regarded as being refiltered by the next block. However, as the original features are retained in each block, we still call our model as a stacking network.

### 2.2. Motivation

The standard DSN is composed of basic DSN blocks. Each block, as developed in [11], consists of a simple, easy-to-learn multilayer perceptron (MLP) with only one hidden layer. For purely discriminative tasks experiments have shown that DSN, despite its simplicity, performs better than the deep belief network [12]. It is pointed out in [13] that higher block is guaranteed to perform better on the training set than the previous block, because the original input is retained. In contrast to other deep architectures, the DSN does not aim to discover effective transformed features. Due to the simplicity of each block, the DSN is considerably easier to train.

As pointed out above, a variant of feature expansion practice is used in our model. In fact, we could directly get a denoised version of features from the separated speech, which is shown in Fig. 2. As can be seen from the figure, the re-extracted features are closer to the clean ones, and are much more structured than the noisy ones. As there is an intuitive

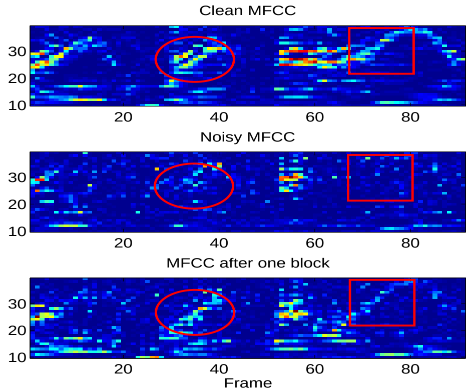

**Fig. 2**. The feature extracted from the separated speech of a DSN block is a denoised version of the noisy features. Top panel: clean MFCC features. Middle panel: corresponding noisy MFCC features at -5 dB mixed with factory noise. Bottom panel: corresponding re-extracted MFCC features.

assumption structured data are easier to map to training targets (another structured data) [14], we could expect better performance could be obtained by the next block of DSN.

The second motivation is as follows. It is verified in [5] that combined features from different domains performs better. We further developed this idea by extracting combined features from combined separated speech in different representative domains, which is named as cross-domain cooperation (CDC) and found to be effective in this study. Besides, multi-task learning can improve the performance of related tasks by joint learning, which has been proven by empirical and theoretical evidences [15, 16]. In a sense, our model is ideologically inspired by and similar to multi-task learning, except that the two tasks here cooperate by providing a cross-domain "input" feature vector rather than a combined "output" label vector.

### 2.3. Feature Extraction and Training Target

As mentioned in Section 1, various features and targets have been investigated with their performance in speech separation tasks. In this study, we use a window of the combined acoustic features proposed in [5] to predict the square-root of ideal ratio mask (IRM) at each time frame.

The combined features include amplitude modulation spectrogram (AMS), relative spectral transformed perceptual linear prediction coefficients (RASTA-PLP), mel-frequency cepstral coefficients (MFCC) and Gammatone filterbank power spectra (GF). All features are smoothed by a second-order ARMA filter [4, 17].

## 3. EXPERIMENTS AND ANALYSIS

### 3.1. Experiment Settings

We use Chinese National Hi-Tech Project 863 sentences, which are recorded by 100 male and 100 female speakers with 500 utterances for each speaker, as the speech corpus. Five broad band noises, i.e., a factory noise, a babble noise, a speech-shaped noise (SSN), a machine noise and a traffic noise are additively mixed with clean speech to create the training and test mixtures. To create the training set, 100 utterances from 5 male and 5 female speakers, with ten utterances for each speaker, are mixed with the first three noises at 0 and -5dB. To create the test set, we randomly choose 50 new utterances from unseen speakers (5 female and 5 male) to mix with each of the five noises at -7, -5, -2 and 0 dB.

In evaluation, we take Short-Time Objective Intelligibility measure (STOI) [18], Perceptual Evaluation of Speech Quality score (PESQ) [19] and SNR as the evaluation metrics. STOI and PESQ can evaluate the objective speech intelligibility and speech quality improvement, respectively. All results are obtained by comparing with the speech signal we obtained through filtering the mixture signal with ideal square-root IRM.

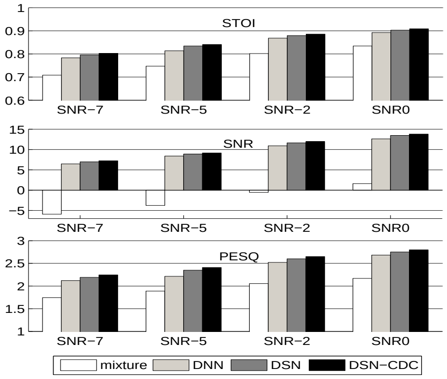

**Fig. 3**. Average STOI, SNR and PESQ results of different models on the test set at different SNRs across five noise types in the cochleagram domain.

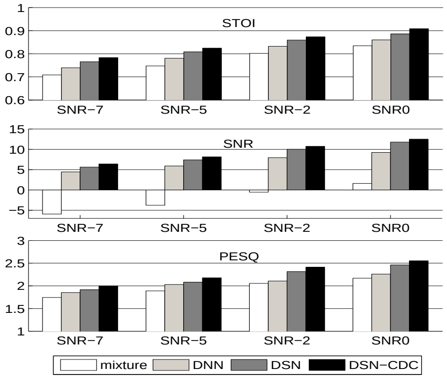

**Fig. 4**. Average STOI, SNR and PESQ results of different models on the test set at different SNRs across five noise types in the spectrogram domain.

To illustrate the effectiveness of our model, we compare it with a standard DNN-based speech separation algorithm [6]. The DNN uses three hidden layers, each having 300 logistic sigmoid units, and bounded (within [0,1]) linear output units. The network is pretrained by RBM with dropout regularization [20]. We use a window (5 frames) of combined features as inputs to the DNN (input dimension is 455). For fair comparison, each block in our DSN-CDC model also has 300 sigmoid hidden units and bounded linear output units.

### 3.2. Result Analysis

Fig. 3 and Fig. 4 show the average results on the test set at different SNRs across five noise types using input features with multiple frames expansion in the two domains. To verify the

**Table 1**. The results that different systems obtains at 0 dB.

| system | matched | | | unmatched | | |
|---|---|---|---|---|---|---|
| | STOI | SNR | PESQ | STOI | SNR | PESQ |
| MIX | 0.84 | 1.34 | 2.15 | 0.85 | 2.18 | 2.25 |
| DNN | 0.92 | 16.44 | 2.97 | 0.88 | 10.90 | 2.54 |
| DSN | 0.93 | 17.95 | 3.09 | 0.89 | 11.24 | 2.57 |
| DSN-CDC | 0.94 | 18.62 | 3.18 | 0.89 | 11.56 | 2.56 |

**Table 2**. The results that different systems obtains at -2 dB.

| system | matched | | | unmatched | | |
|---|---|---|---|---|---|---|
| | STOI | SNR | PESQ | STOI | SNR | PESQ |
| MIX | 0.80 | -0.83 | 2.05 | 0.81 | 0.03 | 2.12 |
| DNN | 0.90 | 15.10 | 2.83 | 0.85 | 8.98 | 2.35 |
| DSN | 0.92 | 16.45 | 2.96 | 0.86 | 9.19 | 2.38 |
| DSN-CDC | 0.93 | 17.02 | 3.01 | 0.87 | 9.37 | 2.39 |

**Table 3**. The results that different systems obtains at -5 dB.

| system | matched | | | unmatched | | |
|---|---|---|---|---|---|---|
| | STOI | SNR | PESQ | STOI | SNR | PESQ |
| MIX | 0.75 | -4.06 | 1.90 | 0.76 | -3.19 | 1.84 |
| DNN | 0.87 | 12.88 | 2.60 | 0.77 | 6.32 | 1.85 |
| DSN | 0.89 | 14.01 | 2.72 | 0.81 | 6.30 | 2.10 |
| DSN-CDC | 0.90 | 14.50 | 2.79 | 0.81 | 6.34 | 2.12 |

**Table 4**. The results that different systems obtains at -7 dB.

| system | matched | | | unmatched | | |
|---|---|---|---|---|---|---|
| | STOI | SNR | PESQ | STOI | SNR | PESQ |
| MIX | 0.72 | -6.21 | 1.76 | 0.72 | -5.32 | 1.77 |
| DNN | 0.84 | 11.20 | 2.42 | 0.75 | 4.27 | 1.90 |
| DSN | 0.86 | 12.17 | 2.52 | 0.76 | 4.43 | 1.95 |
| DSN-CDC | 0.87 | 12.55 | 2.60 | 0.76 | 4.41 | 1.94 |

effectiveness of stacking and cross-domain cooperation learning, we carried out separate experiments to evaluate them (respectively corresponding to the results of DSN and DSN-CDC). From the figure, we can see that the results of DSN are better than the DNN based method. This is due to the stacking of features of separated speech from DSN blocks, which is lacked in DNN that only relies on features extracted and transformed from the original mixtures to do separation. As mentioned in Section 2, this is equivalent to using denoised features to do separation, which has been proved to be effective in [2]. It can also be seen from the figure that DSN-CDC leads to better results than DSN in all the three metrics, which proves the effectiveness of the cross-domain cooperation process. Meanwhile, the cooperative learning brings improvements in both domains, especially in the spectrogram domain. The reason is that the baseline performance of our model in the spectrogram domain is not as well as in the cochleagram domain. It indicates that the separation in the domain with lower baseline benefits more from the cooperation.

In the results above, the DSN-CDC uses three blocks. In our experiments, we have verified the hypothesis that performance improves with the increase of blocks. As can be seen from Fig.5, the STOI metric is getting better with the increase of the number of stacked modules, on both matched-noise and unmatched-noise test set. Similar trends have been found with the other two metrics in our experiments.

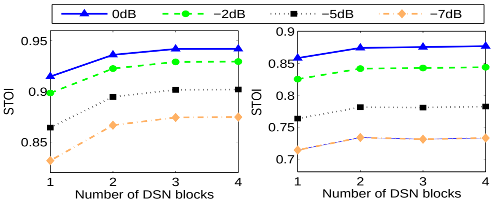

**Fig. 5**. STOI results of the system when more blocks are stacked. Left: matched-noise case. Right: unmatched-noise.

To further investigate the generalization ability, Table 1 to Table 4 show the separation results in matched/unmatched noise conditions at matched/unmatched SNRs in the cochleagram domain. Similar trends appear in the spectrogram domain. The results show that our DSN and DSN-CDC method have better performance and generalize well to unmatched-SNR and unmatched-noise conditions. This indicates that our method has both stronger modeling and generalization ability than the DNN based method. We also note that the DSN-CDC method remarkably outperforms the DSN method, in both matched and unmatched conditions, which indicates the cross-domain cooperation process is effective in obtaining separated speech with higher quality and intelligibility.

## 4. CONCULUDING REMARKS

We have proposed to do speech separation with a cross-domain cooperative deep stacking network (DSN-CDC). The proposed structure can get denoised features from the output of each DSN block, which can greatly improve performance of separation. Besides, experiments have shown that the multi-task learning inspired cross-domain cooperation practice can further boost the separation results. Based on a DSN framework, our structure provides a new feature combination way for cross-domain/task cooperation.

Compared with previous methods, we do not use denoising autoencoders or denoising MLP to get robust transformed features. Thus the supervised learning framework is easy to train, as each block of the DSN has only one hidden layer. However, we have obtained good results using this approach. Experiments show that our method has strong modeling ability as well as generalization ability at very low SNR under matched and unmatched noise conditions.

## 5. REFERENCES

[1] P.C. Loizou, G. Kim, "Reasons why current speech-enhancement algorithms do not improve speech intelligibility and suggested solutions," *IEEE Trans. Audio, Speech, Lang. Process.,* vol. 19, no. 1, pp. 47-56, 2011.

[2] Y.X. Wang, and D.L. Wang, "Feature denoising for speech separation in unknown noisy environments," in *Proc. ICASSP*, 2013.

[3] Y.X. Wang, A. Narayanan, and D.L. Wang, "On training targets for supervised speech separation," Tech. Rep., Technical Report OSU-CISRC-2/14-TR05, Department of Computer Science and Engineering, The Ohio State University, Columbus, Ohio, USA, 2014. Available: ftp://ftp. cse. ohio-state. edu/pub/techreport/2014/TR05. pdf., 2014.

[4] J.T. Chen, Y.X. Wang, and D.L. Wang, "A feature study for classification-based speech separation at very low signal-to-noise ratio," in *Proc. ICASSP*, 2014.

[5] Y.X. Wang, K. Han, and D.L. Wang, "Exploring monaural features for classification-based speech segregation," in *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, pp. 270-279, 2013.

[6] Y. Xu, J. Du, L.R. Dai, C.H. Lee, "An experimental study on speech enhancement based on deep neural networks," in *IEEE Signal Processing Letters*, vol. 21, pp. 65-68, 2014.

[7] S. Nie, H. Zhang, X.L. Zhang, and W.J. Liu, "Deep stacking networks with time series for speech separation," in *Proc. ICASSP*, 2014.

[8] Y. Shao, Z.Z. Jin, D.L. Wang, and S. Srinivasan, "An auditory-based feature for robust speech recognition," in *Proc. ICASSP,* 2009.

[9] P. Vincent, H. Larochelle, Y. Bengio, and P.A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proc. ICML*, 2008, pp. 1096C1103.

[10] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.A. Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," in *The Journal of Machine Learning Research,* vol. 11, pp. 3371-3408, 2010.

[11] L. Deng, and D. Yu, "Deep convex networks: a scalable architecture for speech pattern classification," in *Proc. Interspeech*, 2011.

[12] L. Deng, D. Yu, and J. Platt, "Scalable stacking and learning for building deep architectures," in *Proc. ICASSP*, 2012.

[13] B. Hutchinson, L. Deng, and D. Yu, "Tensor deep stacking networks," in *IEEE Trans. Pattern Analysis and Machine Intelligence (special issue of Learning Deep Architectures)*, 2013.

[14] Y.X. Wang, and D.L. Wang, "A structure-preserving training target for supervised speech separation," in *Proc. ICASSP*, 2014.

[15] A. Argyriou, T. Evgeniou, and M. Pontil, "Convex multi-task feature learning," in *Machine Learning*, vol. 73, no. 3, pp. 243-272, 2008.

[16] R. Collobert and J. Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning," in *Proceedings of the 25th international conference on Machine learning*. ACM, 2008, pp. 160-167.

[17] C. Chen and J. A. Bilmes, "MVA processing of speech features," *IEEE Trans. Audio, Speech, Lang. Process.,* vol. 15, no. 1, pp. 257-270, 2007.

[18] C.H. Taal, R.C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," in *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 7, pp. 2125-2136, 2011.

[19] 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 *Proc. ICASSP*, 2001.

[20] G.E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R.R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," *arXiv preprint arXiv: 1207.0580*, 2012.