Multi-resolution Stacking for Speech Separation Based on Boosted DNN

Xiao-Lei Zhang, DeLiang Wang

Department of Computer Science & Engineering and Center for Cognitive & Brain Sciences,
The Ohio State University, Columbus, OH, USA

huoshan6@126.com, dwang@cse.ohio-state.edu

Abstract

Recent progress in speech separation shows that deep neural networks (DNN) based supervised methods can improve the performance in difficult noise conditions and exhibit good generalization to unseen noise scenarios. However, existing approaches do not explore contextual information sufficiently. In this paper, we focus on exploring contextual information using DNN. The proposed method has two parts—a multi-resolution stacking (MRS) framework and a boosted DNN (bDNN) classifier. The MRS framework trains a stack of classifier ensembles, where each classifier in an ensemble concatenates the raw acoustic feature and the outputs of its bottom ensemble as a new feature, and different classifiers in an ensemble work with different window lengths. The bDNN classifier first generates multiple base predictions for a frame from a given window that is centered on the frame and contains multiple neighboring frames, and then aggregates the base predictions for the final prediction. Our experimental comparison with DNN based speech separation in difficult noise scenarios demonstrates the effectiveness of the proposed method in terms of both prediction accuracy and objective speech intelligibility.

Index Terms: boosted deep neural networks, contextual information, multi-resolution stacking, speech separation.

1. Introduction

Speech separation aims to separate speech from its noise mixture. It has various real-world applications, such as hearing-aids, robust speech recognition [1, 2], and speech communications. Speech separation techniques can be roughly categorized to three classes—signal processing based ones, statistical model based ones [3, 4], and recent machine learning based ones [5–8, 8–11, 11–16]. The first two classes are efficient, but they have some limitations in extremely difficult noise environments. Specifically, signal processing based methods, such as Wiener filtering, minimum mean square error, and spectral subtraction, usually make strong assumptions about the interference, e.g. quasi-stationarity, which limits their applications to a general acoustic background. Statistical model based ones [3,4] build various simple statistical models, such as Gaussian models, to model/smooth acoustic features. These methods generally work well when the background noise is relatively stationary, however, they are not very effective in capturing highly variant and non-stationary noise distribution, particularly when the signal-to-noise ratio (SNR) is low.

Machine learning based approaches, either unsupervised [5–8] or supervised [8–11, 11–16], reformulate speech separation to a classification/regression problem. One type of unsupervised learning methods [5] take each time-frequency (T-F) unit as a single data point and assign it to one of the potential speakers by clustering, which is quite computationally costly. Another type of unsupervised methods explore domain-specific assumptions of data [6–8], such as assuming the accurate detection of silence period for estimating the basis vectors of nonnegative matrix factorization [6] or the sparse property of noise distribution for robust principle component analysis [7], which may not be satisfied easily.

Supervised learning based approaches [8–11,11–16], which can incorporate multiple acoustic features, prior knowledge, and contextual information well, become more and more popular. Besides traditional Gaussian mixture models and support vector machines, recent popular supervised methods also include nonnegative matrix factorization [8, 10, 11] and deep neural network (DNN) [11–16]. Particularly, when trained with large-scale data and vast amount of noise types, DNN based methods demonstrated strong generalization ability to unseen noise scenarios [12–16]. Further given the fast on-line prediction of DNN, they are quite promising in industrial applications. Hence, we pursue DNN based methods.

In this paper, we explore contextual information of data by a simple hierarchical framework, named multi-resolution stacking (MRS) (see Section 2.2). We also propose a compressed ensemble of classifiers, named boosted DNN (bDNN) (see Section 2.3), as the base classifier of the MRS framework. The experimental results in Section 5 show that in terms of both prediction accuracy and objective speech intelligibility, (i) the proposed method outperforms DNN based speech separation when they are given the same amount of training data; (ii) the proposed method has a strong generalization ability that even a small amount of training data can make it as effective as the DNN based method with a large amount of training data.

2. Algorithm description

2.1. Preliminary

Common training targets of supervised speech separation (i.e. ground-truth label) include clean spectrum, ideal binary mask (IBM), and ideal ratio mask [14]. In this paper, we use IBM [17] as our computational target. It is a time-frequency binary mask defined as:

$$y_{m,f} = \begin{cases} 1 & \text{if } \ SNR(m,f) \geq LC \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

where m and f index the time and frequency of an audio time-frequency unit, *SNR* represents the ground-truth signal-to-noise

ratio (SNR) at the unit, and *LC* (named local SNR criterion) is a predefined threshold for a given SNR level. The value 1/0 means that the unit is dominated by speech/noise respectively. In this paper, the IBM is generated by first filtering the raw waves of clean speech and its noise mixture by an l-band gammatone filterbank, then calculating the local SNR of each time-frequency unit, and at last conducting masking by equation (1).

We formulate supervised speech separation as a classification problem. Suppose the classifier is trained on $\{(\mathbf{x}_m, \mathbf{y}_m)\}_{m=1}^{M}$ and tested on a different set $\{\mathbf{x}_n\}_{n=1}^{N}$. The input of the classifier is the acoustic features of frames $\{\mathbf{x}_m\}_{m=1}^{M}$. The ground-truth label of $\mathbf{x}_m$ is the IBM at the mth frame, which is an l-dimensional binary vector denoted as $\mathbf{y}_m = [y_{m,1}, \ldots, y_{m,f}, \ldots, y_{m,l}]^T$.

Note that the proposed method in this paper is not limited to the IBM, it can use many other training targets.

2.2. Multi-resolution stacking

It is known that contextual information is important in improving the performance. One common technique to incorporate contextual information is to train models with a fixed window length that performs the best among several choices of window lengths. We denote the technique of adding a window to incorporate neighboring frames the *resolution*. Here, we argue that (i) for a certain task, although only one resolution performs the best, other resolutions may still provide useful information that may further improve the performance; (ii) although we can manage to pick up the best resolution for a certain task, it is still inconvenient to do so case by case. We propose a simple framework, named multi-resolution stacking, to solve the two problems together.

As described in Figure 1, MRS is a stack of classifier ensembles. In the training stage of MRS, suppose we are to train S building blocks. The sth building block has K_s classifiers, denoted as $\{f_{s,k}(\cdot)\}_{k=1}^{K_s}$. The kth classifier $f_{s,k}(\cdot)$ takes $\mathbf{z}_s$ as the input:[1]

$$\mathbf{z}_s = \begin{cases} \mathbf{x} & \text{if} \quad s = 1 \\ \left[\hat{\mathbf{y}}_{s-1,1}^T, \ldots, \hat{\mathbf{y}}_{s-1,K_{s-1}}^T, \mathbf{x}^T\right]^T & \text{if} \quad s > 1 \end{cases} \tag{2}$$

and takes $\mathbf{y}$ as the training target, where $\{\hat{\mathbf{y}}_{s-1,k'}\}_{k'=1}^{K_{s-1}}$ are the soft predictions of the training frame $\mathbf{x}$ produced by the $(S-1)$th building block. After $f_{s,k}(\cdot)$ is trained, it produces a soft prediction $\hat{\mathbf{y}}_{s,k}$ of $\mathbf{z}_s$ for the upper building block.

If the resolution of $f_k(\cdot)$ is $W_{s,k}$ $(k = 1, \ldots, K_s)$, then it uses the extended feature

$$\mathbf{z}'_s = \left[\mathbf{z}_{s,-W_{s,k}}^T, \mathbf{z}_{s,-W_{s,k}+1}^T \ldots, \mathbf{z}_{s,0}^T, \ldots, \mathbf{z}_{s,W_{s,k}-1}^T, \mathbf{z}_{s,W_{s,k}}^T\right]^T \tag{3}$$

instead of the original feature $\mathbf{z}_s$ as its training feature, where the subscript 0 in $\mathbf{z}_{s,0}$ is a general index for describing any training frame.

Doubling resolution W will double the size of training data. Hence, MRS is hard to handle both a large W and a large training set. To reduce the memory requirement of computing power, we present a trick: one can pick a subset of frames within the window instead of all frames. In this paper, we pick the frames indexed by $\{-W, -W+u, -W+2u, \ldots, -1-u, -1, 0, 1, 1+u, \ldots, W-2u, W-u, W\}$, where u is a user defined integer parameter. This trick not only makes all classifiers in a building block have the same amount of memory requirement but also does not decrease the performance significantly in experience.

[1]For clarity, we omit the time index of the training and test frames in this subsection.

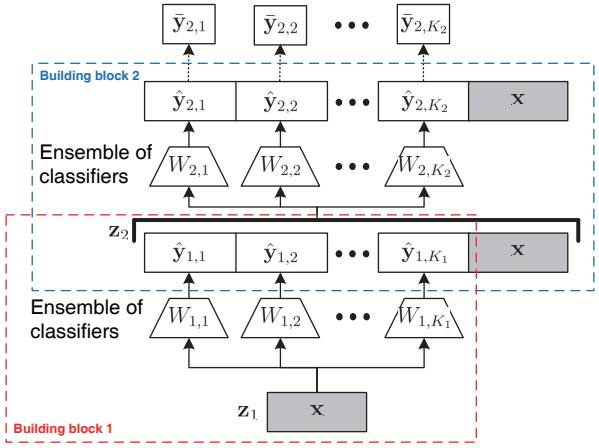

Figure 1: Diagram of the multi-resolution stacking training. The variables in the figure are defined in Section 2.2. The diagram draws two building blocks. Each trapezoid module represents a base classifier.

In the test stage of MRS, we get a serial soft predictions as we did in the training stage from bottom up. Different from the training stage, the test stage has a hard decision step: after getting the output of the Sth building block, we do a hard decision on the output of any classifier in the building block, e.g. $\hat{\mathbf{y}}_{S,k}$, by:

$$\bar{y}_{S,k,f} = \begin{cases} 1 & \text{if} \quad \hat{y}_{S,k,f} \geq \delta \\ 0 & \text{otherwise} \end{cases}, \quad \forall f = 1, \ldots, l \tag{4}$$

and take $\bar{\mathbf{y}}_{S,k}$ as the final prediction, where $\hat{y}_{S,k,f}$ (or $\bar{y}_{S,k,f}$) is the fth element of $\hat{\mathbf{y}}_{S,k}$ (or $\bar{\mathbf{y}}_{S,k}$), and δ is a decision threshold tuned on a development set.

As will be shown in the experiment, when we increase the number of building blocks, the prediction accuracy will be improved, and the variation between different classifiers will be reduced.

2.3. Boosted DNN for speech separation

In this section, we introduce a classifier, named boosted DNN, as the base classifier of MRS. The idea is motivated from ensemble learning, which first produces multiple different predictions and then averages them for a better prediction. Ensemble learning is quite useful in practice, but it is too costly, particularly for large-scale models, such as neural networks and decision trees. Here we combine the ideas of ensemble learning and model compression by generating multiple different predictions from a single DNN.

Given the mth frame $\mathbf{z}'_m$ defined in equation (3), bDNN trains a standard DNN with a modified training target $\mathbf{y}'_m = [\mathbf{y}_{m-W}^T, \mathbf{y}_{m-W+1}^T, \ldots, \mathbf{y}_m^T, \ldots, \mathbf{y}_{m+W-1}^T, \mathbf{y}_{m+W}^T]^T$, where W is the resolution of bDNN.

In the test stage of bDNN, after getting the prediction result $\hat{\mathbf{y}}'_n$ of $\mathbf{z}'_n$, i.e. $\hat{\mathbf{y}}'_n = \left[\hat{\mathbf{y}}_{n-W}^{(-W)T}, \ldots, \hat{\mathbf{y}}_n^{(0)T}, \ldots, \hat{\mathbf{y}}_{n+W}^{(W)T}\right]^T$, we

extract multiple *base predictions* of $\mathbf{z}'_n$ from $\{\widehat{\mathbf{y}}'_{n+w}\}_{w=-W}^{W}$, denoted as $\left\{\hat{\mathbf{y}}_{n+w}^{(w)}\right\}_{w=-W}^{W}$ accordingly, where $\widehat{\mathbf{y}}'_{n+w} = \left[\hat{\mathbf{y}}_{n+w-W}^{(-W)\,T}, \ldots, \hat{\mathbf{y}}_{n+w}^{(0)\,T}, \ldots, \hat{\mathbf{y}}_{n+w+W}^{(W)\,T}\right]^T$ represents the output prediction of a frame by DNN that is w frames behind $\mathbf{z}'_n$, and $\hat{\mathbf{y}}_{n+b}^{(a)}$ represents the ath base prediction of the frame $\mathbf{z}_n$ which is contained as part of the output prediction of $\mathbf{z}'_{n+b}$ from DNN. Then, we aggregate the base predictions as follows:

$$\hat{y}_{n,f} = \frac{\sum_{w=-W}^{W} \hat{y}_{n,f}^{(w)}}{2W+1}, \quad \forall f = 1, \ldots, l. \tag{5}$$

Finally, the predicted binary mask of $\mathbf{z}_n$ is calculated by equation (4).

The key advantage of bDNN is that it not only can generate multiple diverse predictions but also reduces the time and storage complexities to that of training a single DNN model. Its weakness is that because all different predictions share the same DNN generator, the diversity between the predictions is not as large as the situation that all predictions are generated independently from multiple DNNs.

3. Feature description

Multi-resolution cochleagram feature, first proposed in [18], is used as the acoustic feature. For the integrality of this paper, we present it briefly as follows. Multi-resolution cochleagram is a concatenation of 4 cochleagram features with different window sizes and different frame lengths. The first and fourth cochleagram features are the 64-dimensional log-scale energy of frames with frame lengths set to 20 ms and 200 ms respectively, where each dimension is calculated from one corresponding channel of a 64-channel gammatone filterbank. The second and third cochleagram features are calculated by smoothing each time-frequency unit of the first cochleagram feature with two square windows that are centered on the unit and have the sizes of 11×11 and 23×23. After calculating the 256-dimensional multi-resolution cochleagram feature, we calculate its Deltas and double Deltas, and then combine all three into a 768-dimensional feature.

4. Relationship to prior work

We summarize the relationship of this study to previous work as follows. (i) The stacking training is motivated from the superior generalization ability of stacked generalization [19] and more recent deep tensor stacking networks [20, 21] over their building blocks. Different from [19], our method takes the raw feature as part of the input of each building block. Different from [20, 21], each building block of our method is a bDNN ensemble but not a single neural network with only one hidden layer, and also, each block of our method takes the output of its direct predecessor block as its input instead of taking the outputs of all predecessor blocks into account. (ii) The multi-resolution scheme is motivated from the effectiveness of multi-resolution cochleagram feature [18] over its components. Different from [18], our method uses the training method to incorporate contextual information instead of simply averaging the amplitude of neighboring units, so that it tends to be more powerful in fusing contextual information. (iii) bDNN is generalized from [22] and is rooted in ensemble learning. Our method can be categorized to the ensemble methods of manipulating training features [23]. However, different from common ensemble methods which train a lot of weak learners with costly computational budgets, we can get a set of predictions from only one DNN. (iv) bDNN is also related to the model compression of ensemble methods [24, 25]. Different from [24, 25], bDNN compresses an ensemble of models by extending the target of a single DNN, while the methods in [24, 25] compress an ensemble of models by first pretraining an ensemble and then using the soft output of the ensemble as the target of a new DNN.

Our method is also different from existing speech separation techniques in [8, 10, 11, 13, 14] which take neither multi-resolution training nor stacking. They also did not explore contextual information as heavily as ours.

5. Experiments

5.1. Experimental settings

We used the clean speech corpus of AURORA4 [26] corrupted by the 'babble" and "factory" noise in the NOISEX-92 noise corpus in extremely low SNR levels (i.e. $[-5, 0, 5]$ dB). We randomly selected 30 and 300 utterances from the clean corpus as our training sets, 20 utterances as our development set, and 60 test utterances for testing. Note that for each noisy corpora, the additive noises for training, development, and test were cut from different intervals of a given noise. For training each DNN/bDNN by backpropagation, we picked the model that achieved the highest area under the receiver operating characteristic curve (AUC) on the development set. We chose the decision threshold δ (in equation (4)) that achieved the highest speech hit rate minus false alarm rate (HIT$-$FA) on the development set as the operating point. We used a 32-band gammatone filterbank to generate the IBM (i.e. $l = 32$).

We compared with the DNN based speech separation that took the raw feature $\mathbf{x}$ as the input and predicted the 32-dimensional IBM directly. Because bDNN uses only one DNN model, we set the DNN models of both the DNN based speech separation and bDNN with the same parameter setting as follows for a fair comparison. The hidden units were tanh function. The output units were sigmoid function. Neither pretraining nor dropout was used. The DNN has two hidden layers. The number of units of each hidden layer was set to 1000. The number of epoches was set to 70. The batch size was set to 512. The learning rate for the adaptive stochastic gradient descent was set to 0.0008. Note that we have tried various parameter settings and found that the aforementioned setting performed best, particularly, using tanh function was consistently better than using sigmoid function or rectified linear unit.

For our MRS training, we trained 2 building blocks (i.e. $S = 2$). Each building block trained 4 bDNNs with resolution parameters (W, u) set to $\{(3, 1), (5, 2), (9, 4), (13, 6)\}$ respectively. We took the hard decision on the soft output of the bDNN with a parameter $(5, 2)$ at the top building block as the final result of MRS. For comparison, we also took the hard decision on the soft output of the bDNN with a parameter $(5, 2)$ at the bottom building block as the performance of the bDNN based speech separation. Note that given the aforementioned parameters, each resolution selected only 7 frames, which means all bDNNs had the same storage complexity.

We used AUC, HIT$-$FA, and short-time objective intelligibility (STOI) [27] as the evaluation metrics, where AUC evaluates the overall quality of the soft prediction $\hat{\mathbf{y}}$ (compared to the IBM), HIT$-$FA evaluates the prediction accuracy at the optimal operating point (i.e. $\bar{\mathbf{y}}$), and STOI evaluates the intelligibility of the time-domain speech signal that is resynthesized from the predicted binary mask. For all metrics, the higher the value is,

Table 1: Performance comparison between the DNN based, bDNN based, and MRS based speech separation methods in three evaluation metrics—AUC, HIT−FA, and STOI, given a training corpus of **30** utterances. The numbers in bold indicate the best results.

		AUC (%)			HIT−FA (%)			STOI		
Noise type	**SNR**	DNN	bDNN	MRS	DNN	bDNN	MRS	DNN	bDNN	MRS
Babble	−5 dB	79.00	81.93	**82.95**	42.33	46.96	**49.49**	0.5632	0.5871	**0.5902**
	0 dB	85.96	88.92	**89.50**	55.30	61.30	**63.13**	0.7118	0.7354	**0.7376**
	5 dB	90.90	93.21	**93.80**	66.04	71.12	**72.98**	0.8247	0.8366	**0.8399**
Factory	−5 dB	79.15	83.56	**84.80**	43.64	50.85	**53.21**	0.5572	0.5861	**0.5871**
	0 dB	87.68	90.75	**91.37**	59.42	65.71	**67.15**	0.7136	0.7375	**0.7385**
	5 dB	92.02	94.37	**95.14**	69.23	74.08	**76.18**	0.8309	0.8455	**0.8515**

Table 2: Performance comparison between the DNN based, bDNN based, and MRS based speech separation methods, given a training corpus of **300** utterances.

		AUC (%)			HIT−FA (%)			STOI		
Noise type	**SNR**	DNN	bDNN	MRS	DNN	bDNN	MRS	DNN	bDNN	MRS
Babble	−5 dB	82.31	83.95	**84.22**	47.88	50.56	**51.30**	0.6034	0.6124	**0.6211**
	0 dB	89.50	91.37	**91.68**	62.74	66.73	**67.81**	0.7480	0.7559	**0.7630**
	5 dB	93.78	94.87	**95.09**	72.91	75.19	**76.26**	0.8384	0.8510	**0.8544**
Factory	−5 dB	84.52	86.89	**87.48**	52.80	57.08	**58.06**	0.5996	0.6133	**0.6176**
	0 dB	91.31	93.28	**93.82**	67.22	71.36	**73.05**	0.7430	0.7606	**0.7654**
	5 dB	94.94	96.34	**96.73**	76.54	79.79	**81.24**	0.8517	0.8623	**0.8643**

the better the performance is.

5.2. Results

Tables 1 and 2 list the performance comparison between the DNN based, bDNN based, and MRS based speech separation methods in three evaluation metrics. From the tables, we observed the following experimental phenomena. (i) The proposed MRS is consistently the best one in all evaluation metrics and noise scenarios, while bDNN performs better than the baseline DNN. (ii) When we do not have enough training data, the superiority of bDNN and MRS over DNN is apparent. (iii) MRS trained with a corpus of only 30 utterances is as effective as DNN trained with a corpus of 300 utterances, which demonstrates the strong generalization ability of MRS. The last two phenomena are important for the practical use of MRS. As shown in [13], the strong generalization of DNN to unseen noise scenarios is usually made by collecting a large number of noise scenarios. If each noise scenario needs only a small fraction of speech, as our MRS did, the scales of training corpora can be significantly reduced.

Table 3 shows the performance standard-deviations of the 4 component outputs of MRS at each building block. From the table, we observed that the standard-deviations are reduced along with the increase of the number of building blocks, which implies that when more building blocks are stacked, the bDNN classifiers in the top building blocks are improved together and tend to be similar.

Table 3: Standard-deviations of the performance of the 4 component outputs of MRS at the bottom (denoted as MRS1) and top (denoted as MRS2) building blocks. The reported variances are average ones over all 6 noise scenarios. The results in the upper half of the table is produced from a training corpus of 30 utterances. The results in the bottom half is produced from a training corpus of 300 utterances.

		AUC (%)	HIT−FA (%)	STOI
30 utterances	MRS1	0.33	0.64	0.0031
	MRS2	0.17	0.40	0.0019
300 utterances	MRS1	0.19	0.37	0.0022
	MRS2	0.10	0.21	0.0011

6. Conclusions

In this paper, we have proposed MRS framework to explore the contextual information of supervised speech separation deeply, and then proposed bDNN as the base classifier of MRS. Specifically, MRS is a stack of classifier ensembles. The classifiers in an ensemble share the same input feature, work with different window lengths (i.e. different resolutions), and generate multiple different predictions which will be further concatenated together with the raw acoustic feature for the upper ensemble. bDNN first produces multiple different base predictions of a given frame by bootstrapping its contextual information, and then averages the base predictions for the final prediction.

Experimental results on extremely difficult environments have shown that the proposed method produces better performance than the DNN based speech separation method, and moreover, the proposed method trained with a small amount of data yields equivalently good performance as the DNN based method with a large amount of data. These results demonstrate the strong generalization ability of the proposed method.

7. Acknowledgements

We thank the Ohio Supercomputing Center for providing computing resources. The research was supported in part by an AFOSR grant (FA9550-12-1-0130).

8. References

[1] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 1, pp. 30–42, 2012.

[2] J. Li, L. Deng, Y. Gong, and R. Haeb-Umbach, "An overview of noise-robust automatic speech recognition," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 4, pp. 745–777, 2014.

[3] R. C. Hendriks, R. Heusdens, and J. Jensen, "MMSE based noise PSD tracking with low complexity," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2010, pp. 4266–4269.

[4] C. Breithaupt and R. Martin, "Analysis of the decision-directed snr estimator for speech enhancement with respect to low-snr and transient conditions," *IEEE Trans. Audio, Speech, Lang., Process.*, vol. 19, no. 2, pp. 277–289, 2011.

[5] F. R. Bach and M. I. Jordan, "Learning spectral clustering, with application to speech separation," *J. Mach. Learn. Res.*, vol. 7, pp. 1963–2001, 2006.

[6] M. N. Schmidt and J. Larsen, "Reduction of non-stationary noise using a non-negative latent variable decomposition," in *IEEE Workshop Mach. Learn. Signal Process.* IEEE, 2008, pp. 486–491.

[7] P.-S. Huang, S. D. Chen, P. Smaragdis, and M. Hasegawa-Johnson, "Singing-voice separation from monaural recordings using robust principal component analysis," in *Proc. Int. Conf. Acoust., Speech, Signal Process.* IEEE, 2012, pp. 57–60.

[8] N. Mohammadiha, P. Smaragdis, and A. Leijon, "Supervised and unsupervised speech enhancement using nonnegative matrix factorization," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 10, pp. 2140–2151, 2013.

[9] P. Smaragdis, "Convolutive speech bases and their application to supervised speech separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 1, pp. 1–12, 2007.

[10] C. Joder, F. Weninger, D. Virette, and B. Schuller, "Integrating noise estimation and factorization-based speech separation: A novel hybrid approach," in *Proc. IEEE Int. Conf. Acoust, Speech, Signal Process.*, 2013, pp. 131–135.

[11] T. G. Kang, K. Kwon, J. W. Shin, and N. S. Kim, "NMF-based speech enhancement incorporating deep neural network," in *Proc. Interspeech*, 2014, pp. 2843–2846.

[12] Y. Wang and D. L. Wang, "Towards scaling up classification-based speech separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 7, pp. 1381–1390, 2013.

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

[14] Y. Wang, A. Narayanan, and D. L. Wang, "On training targets for supervised speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, in press, 2014.

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

[16] Y. Wang, J. Chen, and D. L. Wang, "Deep neural network based supervised speech segregation generalizes to novel noises through large-scale training," Department of Computer Science and Engineering, The Ohio State University, Columbus, Ohio, USA, Tech. Rep. OSU-CISRC-3/15-TR02, 2015.

[17] D. L. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines*, Kluwer Academic, Norwell MA, 2005, pp. 181–197.

[18] J. Chen, Y. Wang, and D. L. Wang, "A feature study for classification-based speech separation at low signal-to-noise ratios," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 12, pp. 1993–2002, 2014.

[19] D. H. Wolpert, "Stacked generalization," *Neural Netw.*, vol. 5, no. 2, pp. 241–259, 1992.

[20] L. Deng, D. Yu, and J. Platt, "Scalable stacking and learning for building deep architectures," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2012, pp. 2133–2136.

[21] B. Hutchinson, L. Deng, and D. Yu, "Tensor deep stacking networks," *IEEE Trans. Patt. Anal. Machine Intell.*, vol. 35, no. 8, pp. 1944–1957, 2013.

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

[23] T. G. Dietterich, "Ensemble methods in machine learning," *Multiple Classifier Sys.*, pp. 1–15, 2000.

[24] C. Buciluǎ, R. Caruana, and A. Niculescu-Mizil, "Model compression," in *Proc. 12th Int. Conf. Knowl. Disc., Data Min.* ACM, 2006, pp. 535–541.

[25] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[26] D. Pearce 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.

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